Review Article

Which Data to Meta-Analyze, and How?

A Specification-Curve and Multiverse-Analysis Approach to Meta-Analysis

Martin Voracek, Michael Kossmeier, and Ulrich S. Tran

Department of Basic Psychological Research and Research Methods, Faculty of Psychology, University of Vienna, Austria

Abstract: Which data to analyze, and how, are fundamental questions of all empirical research. As there are always numerous flexibilities in data-analytic decisions (a "garden of forking paths"), this poses perennial problems to all empirical research. Specification-curve analysis and multiverse analysis have recently been proposed as solutions to these issues. Building on the structural analogies between primary data analysis and meta-analysis, we transform and adapt these approaches to the meta-analytic level, in tandem with combinatorial meta-analysis. We explain the rationale of this idea, suggest descriptive and inferential statistical procedures, as well as graphical displays, provide code for meta-analytic practitioners to generate and use these, and present a fully worked real example from digit ratio (2D:4D) research, totaling 1,592 meta-analytic specifications. Specification-curve and multiverse meta-analysis holds promise to resolve conflicting meta-analyses, contested evidence, controversial empirical literatures, and polarized research, and to mitigate the associated detrimental effects of these phenomena on research progress.

Keywords: combinatorial meta-analysis, digit ratio (2D:4D), graphical display, multiverse analysis, specification-curve analysis

Structural analogies between meta-analysis and the analysis of primary studies in empirical research have been noted since the inception of meta-analytic methods in the late 1970s. In particular, whereas primary studies deal with and analyze a collection of observations (predominantly, data obtained from individual study participants), meta-analyses deal with and analyze a collection of outcomes of primary studies (predominantly, effect sizes extracted from individual studies).

In the wake of the current (2010s) reproducibility debate and method-reform movement in psychological science and other empirical disciplines (Nelson, Simmons, & Simonsohn, 2018), it has increasingly come to attention that there are numerous flexibilities in data-analytic decisions (now interchangeably termed as researcher degrees of freedom, *p*-hacking, or the data-analytic "garden of forking paths"; Gelman & Loken, 2014; Simmons, Nelson, & Simonsohn, 2011). More specifically, it appears that researchers often disagree, or are uncertain, about which individual observations to include (vs. to exclude) in the analysis of an empirical dataset, and further, which data-analytic strategy is fitting. These phenomena are strikingly demonstrated in field experiments of crowdsourced data-analysis, wherein many data-analysts independently from each other tackle the very same, seemingly simple, research question using the very same dataset (Silberzahn et al., 2018).

In similar vein, meta-analyses often are criticized with regard to their study inclusion criteria (i.e., which studies are eligible vs. which are not), and further, which meta-analytic strategy might be appropriate or optimal (beginning from the choice of the effect-size metric, over possible transformations of these, to the type of meta-analytic modeling itself). Apart from that, at least some questionable research practices (such as *p*-hacking) that pervade primary research (Nelson et al., 2018; Simmons et al., 2011) might be less prevalent or likely in meta-analyses.

For empirical primary studies, there have been recent proposals of methodologists to address and resolve these concerns (which data to include, and how to analyze them). Here, we adopt, modify, and apply the framework of these solutions to meta-analysis; illustrate the potential of this approach with a concrete, fully worked practical application example; indicate further such examples in diverse research fields; include appropriate data-visualization techniques; provide software code within the R software environment for practitioners; and discuss the implications of the approach for broader debates surrounding meta-analyses.

Zeitschrift für Psychologie (2019), *227*(1), 64–82
https://doi.org/10.1027/2151-2604/a000357

Specification-Curve and Multiverse-Analysis Approaches to the Analysis of Primary Studies

Simonsohn, Simmons, and Nelson (2015) noted that, in the data analysis of primary studies, researchers, perhaps oftentimes, disagree on which data points to include, and further, disagree on which statistical tests to calculate. Briefly, these considerations boil down to the fundamental questions of *which* data to analyze, and *how* to analyze them. Simonsohn et al. (2015) foremost illustrated these points by discussing a highly publicized, controversial paper (Jung, Shavitt, Viswanathan, & Hilbe, 2014a), which claimed to show that the perceived femininity (vs. masculinity) of (arbitrarily chosen) names for hurricanes was associated with higher (vs. lower) death toll of these hurricanes in the USA.

The paper appeared as a full report in the prestigious *Proceedings of the National Academy of Sciences of the USA* (*PNAS*) and was highly publicized in diverse news outlets, online social media, and the like, as indicated by an Altmetric attention score of 2,334 (as of end of August 2018; see https://www.altmetric.com/details/2397628#score). To put such an exceedingly high media response as this one in appropriate context, among the 11.68 million research outputs so far indexed by Altmetric, the Jung et al. (2014a) paper ranks 344 (or, at the percentile 99.997).

At the same time, the Jung et al. (2014a) paper triggered a daisy chain of critical, published letters to the editor (Bakkensen & Larson, 2014; Christensen & Christensen, 2014; Maley, 2014; Malter, 2014), along with published author replies and rebuttals (Jung et al., 2014b, 2014c, 2014d), all offering discrepant, and seemingly irreconcilable, views on *which* hurricane data to include and *how* to analyze them. Simonsohn et al. (2015) assembled all these views (or, alternative specifications for data analysis) combinatorially, showed that this yielded 1,728 different ways to analyze more or less the same underlying data set, and further showed that the specific finding of female-named hurricanes being deadlier, as reported in Jung et al. (2014a), belonged to a small subset of analyses (37 out of a total of 1,728 specifications, or 2.1%) which yielded a nominally significant result. Hence, the published main finding of the hurricane paper clearly was not supported. As a side note, a later, independent replication attempt (Smith, 2016), published in a specialist journal and utilizing a much broader data set, also found no support for the main finding of the hurricane paper.

Simonsohn et al. (2015) denominated their approach specification-curve analysis, comprised of the following steps: (1) identification of the (reasonable) specifications for analysis (which data to analyze, and how); (2) combinatorial assembly of these specifications (statistically analyzing all of these); (3) visualization of the different results emerging; (4) inferential statistical procedures (permutation/randomization tests or bootstrap techniques, dependent on the data structure and type of research hypothesis), in order to test whether the results as a whole deviate from the null hypothesis. Most recently, specification-curve analysis has successfully been applied for clarifying the role of birth-order effects in personality traits and cognitive abilities, a line of inquiry which hitherto has produced notoriously inconsistent findings (Rohrer, Egloff, & Schmukle, 2017).

A quite similar proposal was made by Steegen, Tuerlinckx, Gelman, and Vanpaemel (2016), who called their approach multiverse analysis. Comparing multiverse analysis with specification-curve analysis shows that the above steps (1) and (2) are identical, that multiverse analysis proposes different graphical displays for step (3) (a histogram, and an additional tile plot, of the p values, as they emerge from multiverse analysis) than specification-curve analysis does (a specification-curve plot), and that multiverse analysis lacks the inferential statistics of step (4).

These approaches are not without forerunners and congenial ideas, which are rooted in the robustness analysis practices in economics and more generally in the predictor-selection problem in regression analysis. Such similar approaches have recently been advocated in sociology (multimodel analysis: Young, 2018; Young & Holsteen, 2017) and epidemiology (vibration-of-effect analysis: Patel, Burford, & Ioannidis, 2015). Generally, these approaches formally appear less worked-out than the above ones and proceed more incremental than systematical or fully combinatorial. That is to say, available control variables (covariates or confounders) are added step by step to a model to test their influence. Owing to this narrower design and intention, we do not discuss them here further.

Research Synthesis of All Possible Study Subsets: Combinatorial Meta-Analysis

One meta-analytic idea somewhat akin to the specification-curve and multiverse analysis approaches is combinatorial meta-analysis (Olkin, Dahabreh, & Trikalinos, 2012). Mainstream sensitivity analysis in meta-analysis can be viewed as similar to regression diagnostics, in that it follows the leave-one-out method (i.e., of k studies leaving out one study at a time, and recalculating the statistic of interest based on the remaining $k - 1$ studies in the meta-analysis). In contrast, combinatorial meta-analysis calculates the statistic of interest for all possible subsets of studies in the meta-analysis (of which there are $2^k - 1$ subsets, when there are k studies). In addition, to visualize the results of combinatorial meta-analysis (in particular, cross-study effect heterogeneity, depending on the selected study subset included in one meta-analytic scenario of the combinatorial meta-analytic universe), Olkin et al. (2012) proposed a novel meta-analytic graphical display, namely the GOSH (graphical display of study heterogeneity) plot.

Although combinatorial meta-analysis is an elegant and excellent means to identify influential studies in a meta-analysis, as of yet this approach has rarely been used. Further, it quickly becomes computationally infeasible (due to the combinatorial explosion inherent in the term $2^k - 1$) with an increasing number of primary studies desired to include in a meta-analysis.

A Specification-Curve and Multiverse-Analysis Approach to Meta-Analysis

Our proposal is straightforward and simple: briefly, we suggest to adopt, transform, and blend the specification-curve and multiverse analysis approaches, which were developed for the analysis of primary studies, to a specification-curve and multiverse approach to meta-analysis (see Taylor & Munafò, 2016, for a recent call for method triangulation of meta-analytic evidence). This includes adaptations of the inferential statistical test (specifically, a parametric bootstrap procedure) of specification-curve analysis, as well as adaptations of the graphical displays of both specification-curve analysis (descriptive and inferential statistical specification-curve plots) and multiverse analysis (histograms and tile plots of *p* values for all specifications) to the meta-analytic framework. We supply software code for these data visualizations and all respective analyses (https://osf.io/nkv46).

Also central to the context considered here is that, in essence, combinatorial meta-analysis is a brute-force method which simply automatically (and thus quasi blindfold) tests all possible study subsets in one meta-analysis. However, the vast majority of these conceivable subsets, which theoretically can be thought of, would not be regarded as reasonable alternative specifications vis-à-vis study eligibility in any meta-analysis. In that regard, the specification-curve and multiverse approach to meta-analysis can be viewed as a theoretically and conceptually guided, and thus parsimonious, minimal variant of combinatorial meta-analysis. A further important difference is that combinatorial meta-analysis analyzes all study subsets with the same meta-analytic technique, whereas the specification-curve and multiverse meta-analytic approach introduced here allows for several ones (e.g., fixed-effect vs. random-effects modeling). Bearing these differences in mind, we suggest to apply combinatorial meta-analysis in tandem with the conceptually more refined approach introduced here.

Worked Example: Meta-Analytic Specification-Curve and Multiverse Analysis of the Effect of Androgen Receptor Gene CAG Repeat Polymorphisms on Digit Ratio (2D:4D)

Explanatory Background

Our fully worked example is taken from a real, and contested, line of inquiry. In particular, it updates and expands two extant meta-analyses (Hönekopp, 2013; Voracek, 2014) on the same topic with new data, and for the first time utilizes a specification-curve and multiverse analysis approach of meta-analysis. In the following, we provide necessary background information about the research field underlying our example, explain the reasons for selecting this research example, and illustrate why we think that the meta-analytic specification-curve and multiverse analysis approach, along with combinatorial meta-analysis, is informative and insightful with regards to research constellations similar to this one.

Diverse strands of animal research, as accumulated since the late 1950s, suggest that prenatal androgen action (PAA; foremost, testosterone levels, and exposure) have long-lasting, permanent (i.e., so-called organizational, or organizing) effects on the brain, behavioral traits, and disease susceptibility postnatally (Berenbaum & Beltz, 2011; Hines, 2010, 2011). This phenomenon is denominated as prenatal programing and, for the above reasons, of interest for a wide array of research fields (including biological, clinical, developmental, differential, economic, health, personality, and sport psychology).

However, there are obvious barriers to study such effects in humans and in psychological science. For one thing, animal endocrine systems and routes, and the effects of these, may not be directly translatable to humans. On the other hand, prenatal hormone measurement is intractable for human research; human sex-hormonal experimentation (e.g., manipulating embryonic testosterone levels) for ethical reasons is infeasible; and experiments of nature (as provided by early-onset endocrine disorders in humans, such as congenital adrenal hyperplasia, complete androgen insufficiency syndrome, or polycystic ovary syndrome) have their own limitations of insight. Hence, having valid retrospective markers for PAA (i.e., endocrine-sensitive endpoints which are observable and measurable) would be of great value for progress on this nexus of research questions and thus are a research desideratum (Cohen-Bendahan, van de Beek, & Berenbaum, 2005; Voracek, 2011).

Of all such proposed PAA markers proposed over the past few decades (e.g., age at menarche, anogenital distance, finger-ridge count, otoacoustic emissions, and twin-type comparisons of same-sex vs. other-sex dizygotic twin pairs), the second-to-fourth digit ratio (2D:4D) by far is the most frequently investigated one. 2D:4D is a finger-length ratio, namely the length of the index finger (2D), relative to the length of the ring finger (4D). On average, men show lower (smaller) 2D:4D than women. This sex effect is of small-to-medium size ($d < 0.50$; Hönekopp & Watson, 2010). From embryologic studies, it is known that these sex differences and individual differences in 2D:4D emerge early, namely already prenatally, during the testosterone peak occurring after one third of gestational length, which in turn gives rise to sexual differentiation and masculinization of the brain and other tissues. Many contributors to the 2D:4D literature believe that sex and individual differences in 2D:4D are developmentally sufficiently stable, as to ensure that 2D:4D can indeed be taken as the long-desired retrospective PAA marker.

The popularity of the 2D:4D marker for research is indicated by the fact that about one decade after the initiation of this line of inquiry (Manning, Scutt, Wilson, & Lewis-Jones, 1998), according to a scientometric analysis of 2D:4D research, the literature totaled more than 300 published journal reports (Voracek & Loibl, 2009). Using the same literature search strategies as this scientometric account to keep track with the growth of this literature, our current estimate (as of end of August 2018) of the size of the 2D:4D literature amounts to more than 1,400 published journal reports, along with more than 150 published journal abstracts and about 300 unpublished academic theses. While these surely are formidable numbers, after 20 years of research, questions of validity (or, lack thereof) of the 2D:4D marker still permeate the literature. This is mainly due to an apparently widespread lack of replicability of initial 2D:4D research findings by subsequent large-scale investigations and corresponding meta-analyses (e.g., Voracek, Kaden, Kossmeier, Pietschnig, & Tran, 2018; Voracek, Pietschnig, Nader, & Stieger, 2011; Voracek, Tran, & Dressler, 2010). In this sense, 2D:4D research may well be characterized as a contested, if not polarized (Hofmann, 2018), field of investigation. As for its perceived importance to, and popularity in, psychological science, we note that published 2D:4D research preponderantly is conducted at psychology departments, and the journals most frequently publishing 2D:4D papers also are from psychology (Voracek & Loibl, 2009). Further, existing journal special issues on 2D:4D research have appeared in psychology journals (Hennig & Rammsayer, 2007; Voracek, 2011).

Our worked example deals with one central validity claim of the 2D:4D marker, namely, its postulated association with a functional length-variant polymorphism found in the human androgen receptor (AR) gene (i.e., gene variants characterized through varying repetitive patterns, which variations alter the function of the gene). This association has been characterized as the "strongest evidence that androgens affect digit ratio" (Breedlove, 2010, p. 4117). Exon 1 of the human AR gene codes for an amino acid tract, in the form of CAG (polyglutamine) stretches of variable length. These repeat-length polymorphisms vary interindividually and, of particular importance, mediate the efficacy of testosterone action, such that longer CAG stretches are less efficacious, whereas shorter CAG stretches are more efficacious. Various research has found that, within physiologic limits, these CAG effects are linear. The genetically based differential efficacy existing in the human AR is therefore expected to correlate *positively* with 2D:4D, to the extent that the latter reflects testosterone sensitivity. That is, a shorter (and more efficacious) CAG repeat number should correspond to lower (masculinized, or male-typed) 2D:4D, whereas longer (and less efficacious) CAG repeats should correspond to higher (feminized, or female-typed) 2D:4D.

This is what has been observed in the first suchlike study (Manning, Bundred, Newton, & Flanagan, 2003). Despite being based on a small sample ($N = 50$), numerous failures to replicate its findings in subsequent reports, which partly were based on much larger samples, and two meta-analyses of the cumulative empirical evidence, which both yielded null findings (Hönekopp, 2013; Voracek, 2014), the Manning et al. (2003) paper is one of the most-cited 2D:4D publications (Voracek & Loibl, 2009), with about 370 citations in Google Scholar (as of end of August 2018). Of these citations, more than one third (about 150) have accrued *after* the appearance of the two meta-analyses summarizing this literature. In contrast, citation counts for the two meta-analyses in the same database are comparatively low (25 citations each for Hönekopp, 2013, and Voracek, 2014). Further, a citation analysis (Voracek, 2014) of Manning et al. (2003) found that 80% of citations to Manning et al. (2003) cited the report confirmatively (as if there were evidence for 2D:4D/CAG correlations) and 70% cited the report solely (as if there were no further 2D:4D/CAG studies). In addition, citation analyses conducted in the Web of Science database (by citing source and science category) show that the citations garnered by Manning et al. (2003) preponderantly come from psychology journals. In line with this, the top-citing journal is from psychology, as well as four further from the top-10-citing journals of Manning et al. (2003).

Here, we are able to provide an appreciable update of the most recent meta-analysis on this topic (Voracek, 2014) only a few years afterward, because more than a few further 2D:4D/CAG studies have since been published. All of this shows the oftentimes uncertain and limited impact of meta-analyses on their respective literatures, and their sometimes disappointing ability to prevent redundant follow-up research (Habre, Tramèr, Pöpping, & Elia, 2014). Owing to its epistemological scope, a specification-curve/multiverse meta-analysis should be more difficult to ignore, or wiped off, than a further conventional (single-specification) meta-analysis, and may as well safeguard against subsequent, largely overlapping and thus redundant (Ioannidis, 2016; Naudet, Schuit, & Ioannidis, 2017), conventional meta-analyses. This is why we opted to select this research question as the worked example.

Methods

Literature Search for the Meta-Analytic Update

For our worked example (for study details and findings, see Table 1), we update the most recent, and largest, published meta-analysis of CAG effects on 2D:4D (Voracek, 2014), that encompassed 13 studies published up to 2014 (Butovskaya et al., 2012; De Naeyer et al., 2014; Durdiaková et al., 2013; Folland et al., 2012; Hampson & Sankar, 2012; Hurd, Vaillancourt, & Dinsdale, 2011; Knickmeyer, Woolson, Hamer, Konneker, & Gilmore, 2011; Kubranská et al., 2014; Latourelle, Elwess, & Elwess, 2008; Loehlin, Medland, & Martin, 2012; Manning et al., 2003; Mas et al., 2009; Zhang et al., 2013), which provided a maximum of 18 samples (total N = 2,909) for meta-analytic inclusion, originating from nine countries located on five continents (Australia, Belgium, Canada, China, Slovakia, Spain, Tanzania, UK, and USA). Using the same multi-pronged literature search and data retrieval strategies and the same eligibility criteria as in the previous meta-analysis (see Voracek, 2014, for details), we ascertained seven further, more recent, studies (Babková Durdiaková et al., 2017; Chang et al., 2015, Cheng, Zhao, Lu, Liu, & Liu, 2016; Durdiaková, Celec, Laznibatová, Minárik, & Ostatníková, 2016; Durdiaková et al., 2015; Warrington et al., 2018; Zhang et al., 2018), which provided 13 additional samples for inclusion, including two samples from a further country (Denmark: Chang et al., 2015). The updated meta-analysis comprises a maximum of 31 samples, with total N = 10,183. The literature search also detected a duplicate publication (not included in analysis): Zhang et al. (2016), not citing Zhang et al. (2013), analyzed exactly the same sample, and used one half of the data, of the earlier report (by calculating correlations within subgroups defined by the lower/upper quartiles of study variables' distributions).

This corpus of primary studies largely is without author overlap; only one group contributed multiple (albeit relatively small) studies to the meta-analysis (Babková Durdiaková et al., 2017; Durdiaková et al., 2013, 2015, 2016; Kubranská et al., 2014). Apart from the CAG studies, Voracek (2014) also reported 2D:4D meta-analyses for a further AR gene repeat-length polymorphism, namely GGC (also termed GGN, polyglycine) stretches. We skip this further evidence, because the respective literature is much smaller and no additional data have emerged.

Table 1 displays the 2D:4D correlations with CAG repeats for right-hand digit ratio (R2D:4D), as well as for left-hand digit ratio (L2D:4D), and the right-minus-left-hand difference in digit ratio (Δ_{R-L}). Although R2D:4D and L2D:4D are substantially positively correlated, and the Δ_{R-L} difference variable is not independent from its constituents as well, here we follow common conventions of digit ratio research and investigate all three of them separately. Specifically, it has been argued (Hönekopp & Watson, 2010) that R2D:4D shows larger sex differences and stronger, or more reliable, effects with variables of interest than L2D:4D, and that there is directional asymmetry, as well as a sex effect, in Δ_{R-L} (on average, $\Delta_{R-L} < 0$, and more often so, or more pronounced, for men, as compared to women).

The Specification Factors: Which Data to Meta-Analyze, and How

We now turn to the specifications we make for the specification-curve and multiverse meta-analysis of the effects of CAG repeats on 2D:4D. We distinguish between external, or "How" factors (i.e., how to meta-analyze the data), and internal, or "Which" factors (i.e., which data to meta-analyze). We decided to consider two of the former and six of the latter type of factors, as follows.

The first external factor concerns the choice of effect size, because, instead of meta-analyzing Pearson r coefficients, one could opt for transforming these to Fisher's z_r coefficients prior to meta-analysis (as in Voracek, 2014). The second external factor concerns the choice of the meta-analytic model. For instance, whereas Hönekopp (2013) used a random-effects model (REM), Voracek (2014) used the fixed-effect model (FEM). Further, we consider two REM variants, differing in how the between-study variance is estimated, namely the DerSimonian-Laird estimator (DL) and the restricted maximum-likelihood estimator (REML), and an unweighted meta-analytic model (UWM) as well. Although the latter approach clearly is atypical for meta-analysis (wherein the credo is that empirical evidence should be weighted according to its information value, a proxy of which is sample size), it nevertheless is

Table 1. Correlations of 2D:4D with CAG repeats length in the androgen receptor gene: Individual studies and updated meta-analysis

					r		
Study (first author)	Country	Sample	2D:4D measurement	*N*	R2D:4D	L2D:4D	Δ_{R-L}
Manning (2003)	UK	Men	Direct	50	.29*	.005	.36*
Latourelle (2008)	USA	Men	Photocopies	35	.00^{a}	–	–
Latourelle (2008)	USA	Women	Photocopies	72	.00^{a,c}	–	–
Mas (2009)	Spain	Men	Photocopies	72^{b}	–.0685^{b}	–.054^{b}	.002^{b}
Mas (2009)	Spain	Male-to-female transsexuals	Photocopies	63^{b}	.0021^{b}	–.0941^{b}	.1447^{b}
Hurd (2011)	Canada	Men	Digicam photographs	178–180	.006	–.12	.14
Knickmeyer (2011)	USA	Boys	Photocopies	71–74^{b}	.143^{b}	.014^{b}	.108^{b}
Knickmeyer (2011)	USA	Girls	Photocopies	70–74^{b}	–.133^{b,c}	–.028^{b,c}	–.010^{b,c}
Butovskaya (2012)	Tanzania	Men	Direct	103^{b}	.1347^{b}	.1913^{b}	–.0798^{b}
Folland (2012)	UK	Men	Photocopies	71	.10	.20	.00^{a}
Hampson (2012)	Canada	men	Flatbed scans	134	–.085	–.063	–.047
Loehlin (2012)	Australia	Boys	Photocopies	182	–.06	–.13	.10
Loehlin (2012)	Australia	Girls	Photocopies	218	.08^{c}	.14*^{c}	–.06^{c}
Durdiaková (2013)	Slovakia	Boys	Flatbed scans	147	.04	.09	–.085^{b}
Zhang (2013)	China	Men	Photocopies	294	.003	.016	–.022
Zhang (2013)	China	Women	Photocopies	391	.030^{c}	–.018^{c}	.055^{c}
De Naeyer (2014)	Belgium	Men	Direct	677	–.05^{e}	–.03^{e}	–.0275^{b,e}
Kubranská (2014)	Slovakia	Men	Flatbed scans	75	.043	.011	.053^{b}
Chang (2015)	Denmark	Men (Klinefelter syndrome, 47,XXY)	Direct	73	.01^{f}	–	–
Chang (2015)	Denmark	Men (controls)	Direct	73	–.03^{f}	–	–
Durdiaková (2015)	Slovakia	Boys	Flatbed scans	15^{b}	.609*	.312	–
Cheng (2016)	China	Women (premature ovarian failure patients)	Digicam photographs	74	.104^{g}	–.094^{g}	–
Cheng (2016)	China	Women (controls)	Digicam photographs	156	.003^{g}	.076^{g}	–
Durdiaková (2016)	Slovakia	Girls	flatbed scans	51	–.25^{c}	–.28*^{c}	–.005^{c}
Babková Durdiaková (2017)	Slovakia	Men	Flatbed scans	65	≈ .00^{b}	≈ .00^{b}	≈ .00^{b}
Warrington (2018), ALSPAC cohort	UK	Boys	Photocopies	R: 2,615; L: 2,618	.008	–.013	–
Warrington (2018), ALSPAC cohort	UK	Girls	Photocopies	R: 2,718; L: 2,714	.047^{c}	.040^{c}	–

(Continued on next page)

Table 1. (Continued)

					r		
Study (first author)	Country	Sample	2D:4D measurement	*N*	R2D:4D	L2D:4D	Δ_{R-L}
Warrington (2018), QIMR cohort	Australia	Boys	Photocopies	231	−.072	−.135	–
Warrington (2018), QIMR cohort	Australia	Girls	Photocopies	287	.123[c]	.128[c]	–
Zhang (2018)	China	Men	Flatbed scans	336	.06	–	–
Zhang (2018)	China	Women	Flatbed scans	580	−.03	–	–
Samples (total *N*)					31 (10,183)	25 (9,014)	18 (2,912)
Combined *r* [95% CI]					.019 [−.001, .038]	.007 [−.013, .028]	.013 [−.024, .049]
Q (I^2)					35 (14%)	39.7 (40%)	18.3 (7%)

Note. [a]Exact effect size not reported in the original study (but definitely was not nominally significant), and requested additional result details were not received (hence, effect set to zero). [b]Effect size not reported in the original study, or the sample size was further amplified after publication (in either case, supply of the additional results details is gratefully acknowledged). [c]The correlation is for the biallelic mean of CAG repeats. [d]Effect reported merely as "significant" in the original study, and requested additional result details were not received (hence, effect set to just-significant, p = .05, two-tailed). [e]Effect size (β coefficient) estimated from linear mixed-effects model, accounting for dependent data structure (siblings) and adjusted for age, height, and weight. [f]Spearman's r_s. [g]Effect size calculated from t statistic and group sizes, according to dichotomized (short vs. long) CAG repeats. *p < .05 (two-tailed). To ensure analytic reproducibility, effect sizes are not rounded. Datasets for the table are available at https://osf.io/2h73x/ (R2D:4D), https://osf.io/ac96w (L2D:4D), and (https://osf.io/5xud3 (Δ_{R-L}).

interesting because the UWM has similarities with the "cognitive algebra" done in traditional, narrative, unsystematic reviews, namely the attitude of taking evidence "as is", no matter what the respective underlying sample size is. Together, the two How factors make up for 2 × 4 = 8 different ways to meta-analyze the *same* data.

Considering the Manning et al. (2003) study in terms of potentially relevant study features, we notice six of these, which therefore constitute our internal, or Which, factors. Manning et al. (2003) was a study of healthy adult White men, with 2D:4D directly measured from the fingers, and published as a full journal report, with all outcomes relevant for this meta-analysis reported therein. All these six study features (participant sex, age group, group status, ethnicity, 2D:4D measurement method, and publication status) are dichotomous; in theory, these Which factors thus make up for $2^6 = 64$ ways to meta-analyze *different* data subsets. We note that, although specification factors generally are categorical, they are not necessarily confined to dichotomies, such as in this example. Further, there are no missing values on these, as the information either is directly reported in the study or self-evident.

The six study features considered here are topically relevant for the following reasons. Regarding participant sex, analyzing AR gene CAG repeats in women is not as straightforward as it is in men, because the human AR gene is located on the X sex-chromosome, of which men (karyotype 46,XY) have but one, whereas women (46,XX) two, and therefore two AR alleles, of which one per cell is randomly inactivated. 2D:4D/CAG studies involving female samples (see Table 1) therefore use the biallelic mean of CAG repeats for analysis. For this reason, some researchers could object to meta-analyze female samples alongside male samples, or object to consider the evidence from female samples at all. In similar vein, researchers might object to consider non-adult (as opposed to adult samples), patient samples (as opposed to healthy individuals), non-White (as opposed to White samples), and samples with image-based 2D:4D measurement (as opposed to direct measurement), because the original evidence (Manning et al., 2003) was for healthy adult White males, whose fingers were directly measured. Regarding publication status, it is evident that among the primary studies a few only appeared as a published journal abstract, and not as a full report, and further that there also are a few studies for which effect-size guesstimates had to be imputed, because of lack of reporting detail in the published study and nonreceipt of requested additional study results information (Latourelle et al., 2008; Mas et al., 2009; for details, see Table 1). These latter studies have been incorporated in one of the prior meta-analyses (Voracek, 2014), but not in the other one (Hönekopp, 2013). It therefore appears fitting to account for publication status (full report, no guesstimates vs. no full report, and/or guesstimates) as the sixth, and final, of our internal factors. Evidently, this last factor merges several things. This is due to the specifics of the primary literature (see Table 1 note details) and its limited size. For larger meta-analyses, it would be both beneficial and feasible to disentangle these.

As mentioned above, the six study features (our internal factors) give rise to potentially $2^6 = 64$ different study designs. In terms of these study features, the exact antitype of the Manning et al. (2003) study would be a study conducted with a patient sample of non-adult non-White females, with image-based 2D:4D measurement, and not published as a full journal report (and/or involving an effect guesstimate). Unsurprisingly, such an antitype study, with study features maximally dissimilar to those in Manning et al. (2003), does not occur among the known primary studies (Table 1). Rather, the sample most dissimilar to the original report is the female patient sample of Cheng et al. (2016), which still is identical in terms of two study features (adult sample and full report). On the other hand, there are two samples (De Naeyer et al., 2014; Chang et al., 2015: male sample) which are exactly identical on all these six study features to Manning et al. (2003), and the majority of samples is identical for at least four or even five out of the six study features. Table 2 displays the specification matrix of the six internal (or study-feature) factors for the

Table 2. Specification matrix for individual studies accounting for six study feature variables

Study	Participant sex	Age group	Group status	Ethnicity	2D:4D measurement	Publication status
Manning (2003)	X	X	X	X	X	X
De Naeyer (2014)	X	X	X	X	X	X
Chang (2015), men	X	X	X	X	X	X
Hurd (2011)	X	X	X	X		X
Butovskaya (2012)	X	X	X		X	X
Folland (2012)	X	X	X	X		X
Hampson (2012)	X	X	X	X		X
Kubranská (2014)	X	X	X	X		X
Chang (2015), patients	X	X		X	X	X
Babková Durdiaková (2017)	X	X	X	X		X
Latourelle (2008), men	X	X	X	X		
Mas (2009)	X	X	X	X		
Knickmeyer (2011), boys	X		X	X		X
Loehlin (2012), boys	X		X	X		X
Durdiaková (2013)	X		X	X		X
Zhang (2013), men	X	X	X			X
Durdiaková (2015)	X		X	X		X
Warrington (2018), ALSPAC cohort boys	X		X	X		X
Warrington (2018), QIMR cohort boys	X		X	X		X
Zhang (2018), men	X	X	X			X
Latourelle (2008), women		X	X	X		
Mas (2009), patients	X	X		X		
Knickmeyer (2011), girls			X	X		X
Loehlin (2012), girls			X	X		X
Zhang (2013), women		X	X			X
Cheng (2016), controls		X	X			X
Durdiaková (2016)			X	X		X
Warrington (2018), ALSPAC cohort girls			X	X		X
Warrington (2018), QIMR cohort girls			X	X		X
Zhang (2018), women		X	X			X
Cheng (2016), patients		X				X

Note. Studies are ordered in decreasing similarity of study features to the original report of Manning et al. (2003) and, within degree of feature similarity, chronologically and alphabetically. The table entries (X vs. cell left blank) correspond to: male versus female sample (for participant sex), adult versus non-adult sample (for age group), healthy individuals versus patient sample (for group status), White versus non-White sample (for ethnicity), direct versus image-based measurement (for 2D:4D measurement), and published as full journal report and with no effect guesstimates necessary versus any of these (for publication status). The dataset for the table is available at https://osf.io/2h73x/.

primary studies detailed in Table 1, listed by decreasing study-feature similarity to Manning et al. (2003). From this assembly it can also be gleaned that from the theoretical number of $2^6 = 64$ different study designs, only 12 different study designs across 31 retrievable samples so far have been implemented by research.

In accounting for the six internal (Which) factors, we fully combinatorially combined the factor levels, along with the respective superset, across all six factors. That is, the subset of male samples only, the subset of female samples only, and the superset of samples regardless of participant sex (either male, or female) were combined with those of adult samples only, non-adult samples only, and with samples of either age group; in turn, with healthy samples only, patient samples only, and samples of either group status; and so forth across all six factors. This yields $3^6 = 729$ combinations potentially available for analysis. From these combinations, only those containing at least two samples were kept for meta-analysis, and duplicated combinations were not included in analysis. This finally yielded 85, 62, and 52 combinations for the R2D:4D, L2D:4D, and Δ_{R-L} analyses, respectively (or 12%, 9%, and 7% of the theoretically possible total). Each of these 85, 62, and 52 subsets (specified according to the Which, or study-feature, factors) was analyzed according to the 2 (effect-size metric) × 4 (meta-analytic model) = 8 different ways (or How factors) to analyze the same meta-analytic subset, thus yielding a grand total of (8 × 85) + (8 × 62) + (8 × 52) = 1,592 different meta-analytic specifications calculated.

Algorithm for the Combinatorial Meta-Analysis

With a maximum number of 31 available samples (for R2D:4D), well over 2 billion unique subsets ($2^{31} - 1 = 2{,}147{,}483{,}648$ exactly) emerge for a full (exhaustive) combinatorial meta-analysis. While this computationally might still be feasible, it is time-consuming and poses problems with graphically displaying the results due to an abundance of overplotting data points. Conveniently, we chose a random sample of 100,000 different subsets for a combinatorial meta-analysis representative of the full set, using a stratified sampling approach with respect to subset size, such that the most prevalent subset sizes (those of intermediate size) were undersampled, while the rarest subset sizes (those of smallest and of largest size) were oversampled. This was achieved by randomly drawing unique subsets for each possible subset size (one to 31 samples for R2D:4D) separately, until the desired number of 100,000 unique subsets was reached.

Parametric Bootstrap for the Inferential Test of the Specification-Curve Meta-Analysis

To evaluate the descriptive meta-analytic specification-curve plot against the null hypothesis of no effect with an inferential statistical test, we used a parametric bootstrap approach. For each sample from the literature (Table 1), we regarded all study features as fixed, but generated random values as new effect sizes under the assumption that the null hypothesis is true: that is, randomly drawn were values from a normal distribution with an expectation of always zero, but the standard deviation equal to the respective sample's observed standard error (thus corresponding to the FEM of meta-analysis). Then, descriptive specification-curve analysis was applied. This whole procedure was repeated 1,000 times, and the resulting 1,000 bootstrapped specification curves then used to find the respective pointwise 2.5% and 97.5% quantiles as the lower and upper limits for each specification number separately. Exceeding one of these limits would indicate that the actual, descriptive specification curve deviates from the under-the-null scenario of no effect ($r = 0$) with two-tailed testing (in parenthesis, we note that, if desired, one-tailed testing would also be possible).

Open Science Practices

We disclose how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study (Simmons, Nelson, & Simonsohn, 2012). Specifically, as this is a meta-analysis, sample size is not determined, but rather arrived at through literature-search strategies and study inclusion/exclusion criteria (detailed above). The full meta-analytic dataset (see Table 1 note) is accessible via the OSF (Open Science Framework). For the full (i.e., conventional) meta-analytic model (see Results section below and Table 1), we did not exclude any data. Owing to the meta-analytic study format, there were no experimental manipulations. Also, there were no further measures than those appearing in Tables 1 and 2. All statistical manipulations (the How factors) and those due to study inclusion versus exclusion (the Which factors) are detailed above.

The focus here is on method development, and the meta-analysis just an illustrative example; hence, we did not preregister it. However, all components necessary for reproducible data analysis (open data, open materials, and open code) are accessible via the OSF and, because of this repository's characteristics, also comply with the FAIR (findable, accessible, interoperable, re-usable) guiding principles for scientific data (Wilkinson et al., 2016).

Results and Discussion

Table 1 (bottom) contains the results of the updated meta-analysis for the associations of digit ratios (R2D:4D, L2D:4D, and Δ_{R-L}) with CAG repeats. According to these simple fixed-effect meta-analytic summaries, which use Fisher's z_r transformation of the Pearson r coefficients for synthesis, there is no evidence for positive correlations between these variables. All combined effects are very close to zero and have rather tight 95% confidence intervals. Cross-study effect heterogeneity (as indicated by the Q tests and the I^2 values) is relatively low. This updated meta-analysis exactly follows the meta-analytic decisions of Voracek (2014). As such, it is important to note that, as seen through the lens of specification-curve and multiverse meta-analysis, this constitutes not more than a single specification, whereas there are numerous alternative specifications.

Figure 1 (to the left) provides a graphical display corresponding to these summary results (bottom of Table 1). Instead of the classic meta-analytic forest plot, we use an advancement of it, namely the meta-analytic rainforest plot (for details, see Schild & Voracek, 2015; Zhang, Kossmeier, Tran, Voracek, & Zhang, 2017). Figure 1 (to the right) contains the visualization (GOSH plots) of the all-subsets (combinatorial) meta-analyses. As mentioned above, we display random samples of 100,000 meta-analytic subsets, as drawn from the much larger number of possible subsets. The impression from the GOSH plots is straightforward: density estimates of the effect distributions are unimodal (thus not suggestive of influential subsets of studies or individual studies, including Manning et al., 2003, which study is highlighted in these plots) and closely centered around zero (thus not suggesting any real effects). Effect heterogeneity preponderantly is low; except that it is somewhat larger, when Manning et al. (2003) is included, thus indicating that this study really is an outlier.

Figure 2 provides the descriptive meta-analytic specification-curve plots for the three meta-analyses (R2D:4D, L2D:4D, Δ_{R-L}). Whereas Simonsohn et al. (2015), in their corresponding plot for primary data analysis, depicted the regression-model point estimates resulting from the alternative specifications, we display the specifications' summary effects with their associated 95% confidence intervals. Similar to meta-analytic caterpillar plots (i.e., magnitude-sorted forest plots), the summary effects are sorted by magnitude. The number of samples contained in each meta-analytic specification is depicted directly beneath, and the combination of Which and How factors constituting each meta-analytic specification through the area pattern below. This area pattern needs to be contemplated vertically. For facilitating this, we use a spectral-color design, comprised of six spectral colors (ordered from red, orange, yellow, green, blue, to violet), which, *once more*, signifies the number of samples involved in the respective meta-analytic specification (because of this near-redundancy, researchers may, of course, choose which components of this graphical display to keep). The array is such that red, orange, or yellow color (think of hot colors as alarm signals) means that in a given specification only the minimum number of samples is (or small numbers of samples are) involved, whereas violet, blue, or green color (think of cool colors as relaxative) codes for the maximum number (or at least large numbers) of samples involved.

Again, the overall pattern is easy to follow: more or less regardless of the meta-analytic specifications made (in terms of which data are meta-analyzed, and how), no evidence for 2D:4D/CAG associations arises. For R2D:4D, 56 out of 680 specifications (8.2%) yield nominally significant ($p < .05$) positive combined effects, which would support the findings of Manning et al. (2003); for Δ_{R-L}, as few as 4 out of 416 specifications (2.7%); and for L2D:4D, only 6 out of 496 specifications (1.2%). Paralleling the results of Simonsohn et al. (2015) in their specification-curve analysis of the hurricane paper of Jung et al. (2014a), the rate of stray positive results is so low as to be perfectly plausible by chance alone. It is emphasized that it is the dominant pattern (i.e., the majority vote) arising from the space of specifications that counts, not any aggregation of these (e.g., their grand total). The latter neither is intended nor seems justified, as individual specifications partly are appreciably similar and there likely is no "deeper truth" calculable by averaging (Patel et al., 2015).

It is interesting to note that the above small number of nominally significant specifications to a great extent involve rather small meta-analytic subsets (signified through hot colors and long confidence intervals in Figure 2) and tend to surface with UWM analyses. Meta-analysts certainly would not draw inferences from a completely unweighted model which summarizes only a small portion of available studies from the literature. However, at the same time this particular scenario has strong similarities with the actual reasoning and the usual procedures involved in writing stand-alone traditional (narrative, unsystematic) literature reviews, or in drafting the literature review section for the introductory part of an empirical research article. The idiosyncracy inherent in these scenarios is that only a small portion of the totality of research evidence is seen and accounted for, and moreover evaluated in a fashion as if all studies would have identical information value (see Kühberger, Scherndl, Ludwig, & Simon, 2016, for a demonstration of the detrimental effects of this misleading approach). For these reasons, it may well be informative to incorporate UWMs into specification-curve/multiverse meta-analyses on a regular basis. In the case of our worked example, this may also serve to understand the persistence of citations to Manning et al. (2003) in this literature, as well as the neglect of available meta-analyses on the same topic (Hönekopp, 2013; Voracek, 2014).

The inferential meta-analytic specification plots (Figure 3) corroborate the above findings, in that they nowhere deviate from the under-the-null scenario of an underlying zero effect. The slight results differences between these

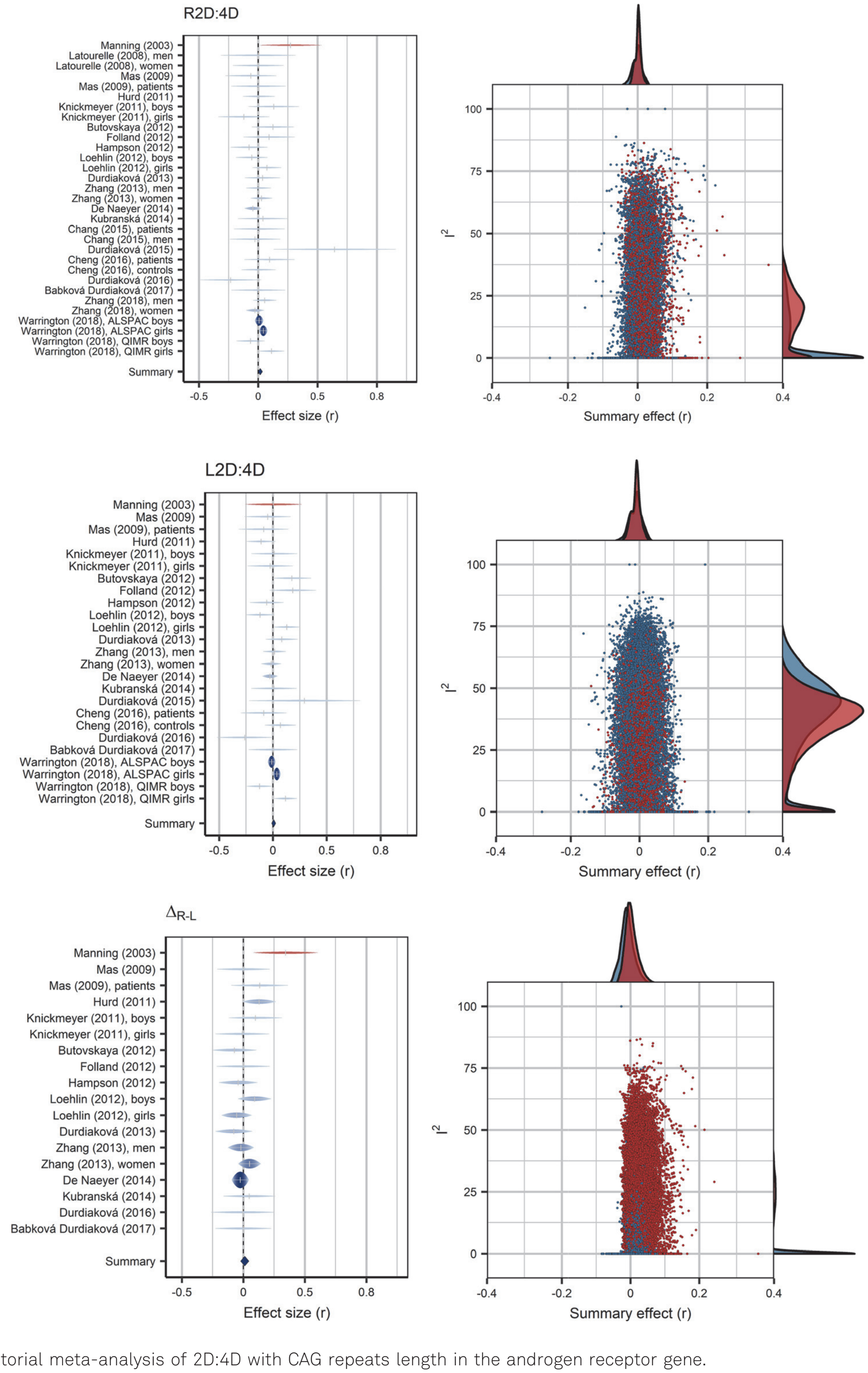

Figure 1. Combinatorial meta-analysis of 2D:4D with CAG repeats length in the androgen receptor gene.

Note. Rainforest plots (on the left; see Schild & Voracek, 2015) for all three meta-analyses visualize study effects as raindrops and the meta-analytic summary effect as diamond at the bottom. Raindrop widths correspond to conventional 95% confidence intervals, while raindrop heights and their shading correspond to the likelihood (i.e., plausibility) of underlying true values, considering the observed study effects, and are proportional to the meta-analytic weight. Study Manning et al. (2003) is highlighted in red. GOSH plots (on the right; see Olkin et al., 2012) show the FEM meta-analytic summary effects on the x axis and the between-study variance statistic I^2 on the y axis for a random sample of 100,000 different study subsets. The distributions of these 100,000 values are visualized by density estimates at the top (for the summary effect) and to the right (for the I^2 values). Study subsets including Manning et al. (2003) are highlighted in red in the color version of this figure available with the online version of the article. R code to reproduce the figure is available at https://osf.io/kqgey/.

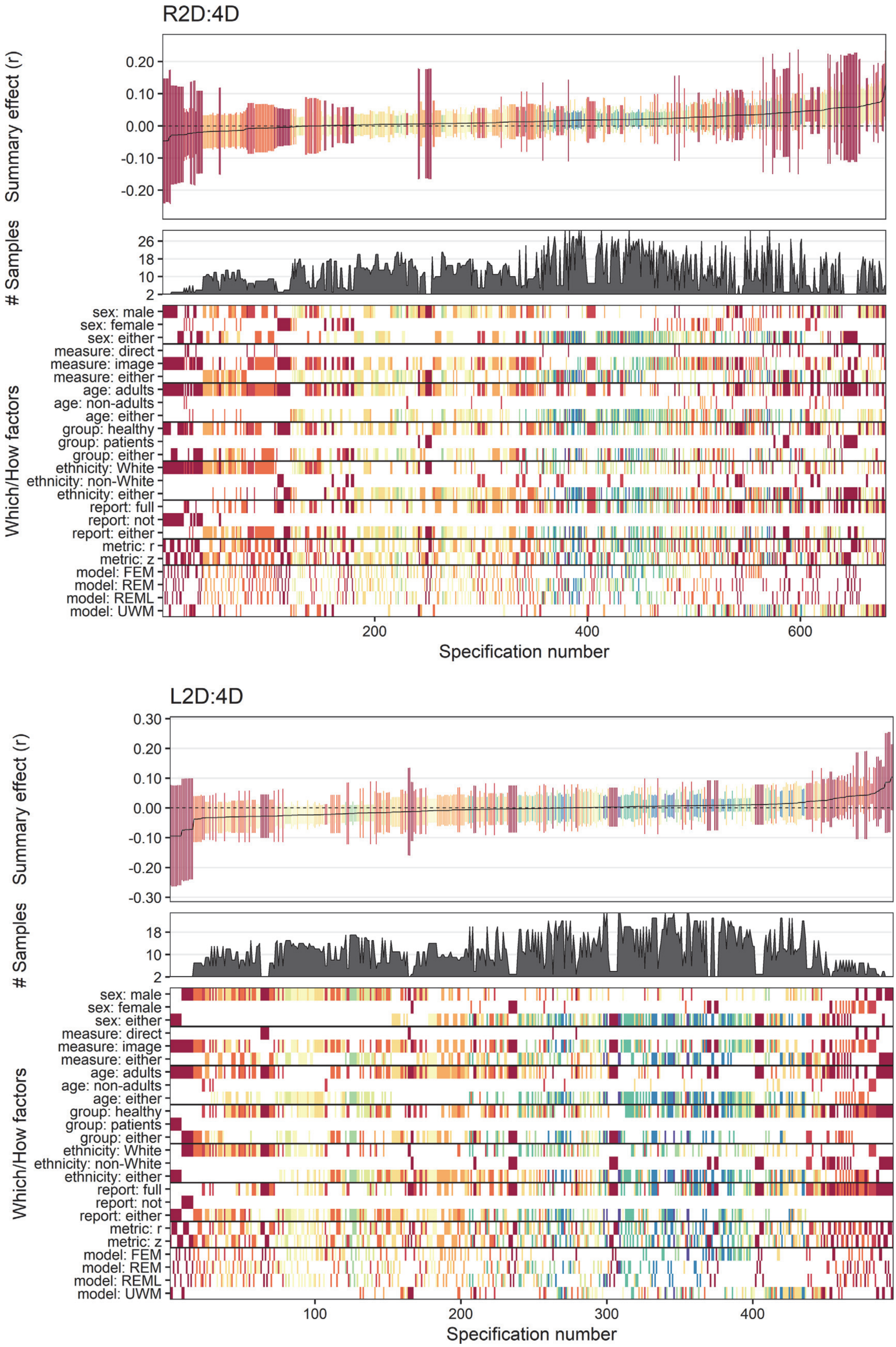

Figure 2. (Continued on next page).

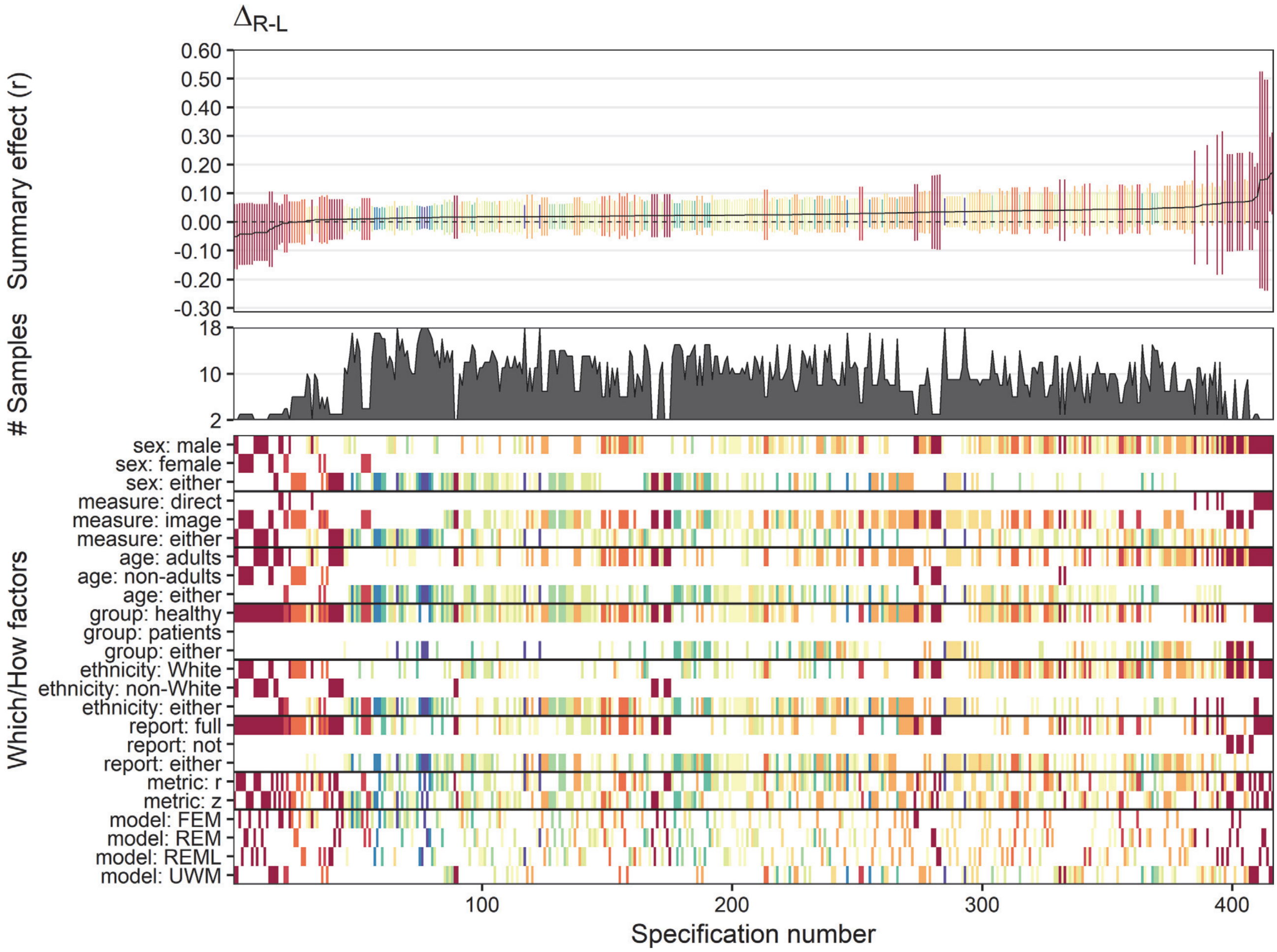

Figure 2. (Continued) Descriptive meta-analytic specification plots for R2D:4D, L2D:4D, and Δ_{R-L}. Descriptive meta-analytic specification plots depict the three specification-curve meta-analyses for R2D:4D, L2D:4D, and Δ_{R-L}. Within each plot, the vertical columns (in the lower half) represent which factor-level combinations of internal (How) and external (Which) specification factors constitute a given specification. In addition, each vertical column is color-coded, signifying the number of samples included in a specification (hot vs. cool spectral colors code for smaller vs. larger number of samples included). The panel in the middle (filled black line chart) likewise shows how many samples are included in a given specification. The top panel shows the resulting meta-analytic summary effects (*r*) for each specification, along with 95% confidence intervals. The summary effects are sorted by their magnitude and connected, resulting in a specification curve. A horizontal dotted line of no effect is inserted at *r* = 0. R code to reproduce the figure is available at https://osf.io/e4bs8/ (R2D:4D), https://osf.io/738sr/ (L2D:4D), https://osf.io/8tw59/ (Δ_{R-L}).

descriptive versus inferential specification plots (some stray positive results vs. none) are understandable through the different evaluation criteria used (null-hypothesis significance testing vs. parametric bootstrap); the conclusions however are identical. Finally, Figure 4 displays histograms of the *p* value distributions for the summary effects of the various meta-analytic specifications (as adopted from the multiverse analysis approach of Steegen et al., 2016). Further conforming with the above evidence of zero effects, no consistent or clear piling up of $p < .05$ values is evident. In principle, the gist of the information provided by these histograms can already be gleaned from the topmost part of the descriptive meta-analytic specification plots (Figure 2). For the sake of completeness, we note that a tile plot of *p* values (constructed similarly to a two-dimensional nested cross-table design, as introduced by Steegen et al., 2016, for multiverse analysis of primary data) would furthermore enable to look up the exact meta-analytic specifications, wherein $p < .05$ values occur. Since we considered hundreds of specifications, the *p* value tile plot would be cluttered and thus is omitted here. Researchers working with fewer meta-analytic specifications might however wish to present a *p* value tile plot in addition (see Steegen et al., 2016, for examples). All in all, although the evidence from our worked example casts a bleak view on the validity status of the 2D:4D marker vis-à-vis genetically based testosterone sensitivity, precisely the exhaustiveness and convergence of this evidence matters and is reassuring.

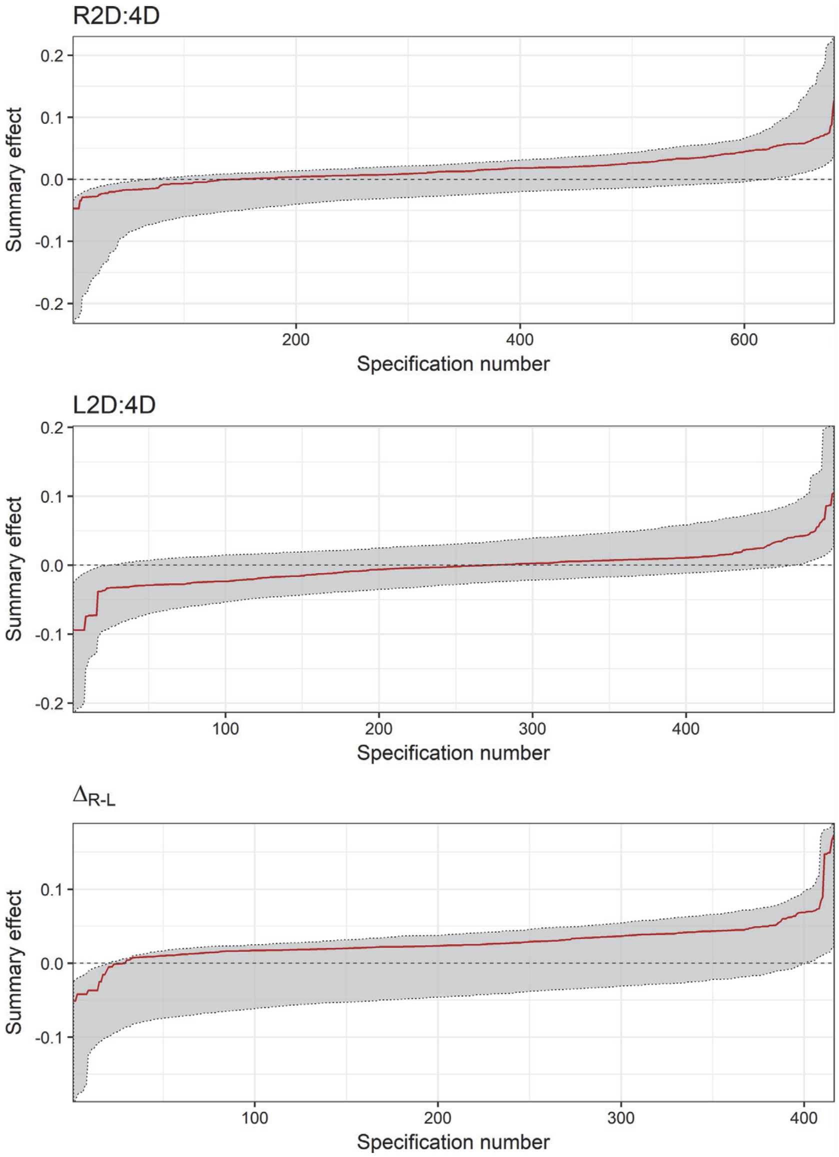

Figure 3. Inferential meta-analytic specification plots for R2D:4D, L2D:4D, and Δ_{R-L}. Inferential meta-analytic specification plots show the specification curve (solid line) of the magnitude-sorted observed meta-analytic summary effects for all specifications. The same curves appear in the corresponding descriptive meta-analytic specification plots (Figure 2). The limits of the gray area correspond to the pointwise 97.5% and 2.5% quantiles of 1,000 specification curves simulated under the null hypothesis for a given specification number, using a parametric bootstrap procedure. Exceeding these limits would constitute evidence against the null hypothesis (r = 0, regardless of specification). R code to reproduce the figure is available at https://osf.io/ru264.

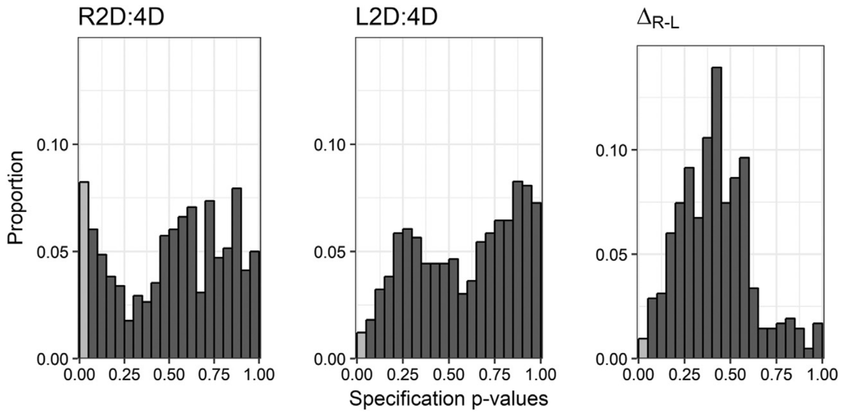

Figure 4. Histograms of the p value distributions for the summary effects of all meta-analytic specifications. Histograms of p values for all meta-analytic specifications, testing whether the meta-analytic summary effect differs from zero (Figure 2). The proportion of nominally significant values ($p < .05$) is in the leftmost column (light gray). R code to reproduce the figure is available at https://osf.io/yu98x.

Conclusions and Implications

We conclude with some further considerations regarding the presented approach. It is important to note that in specification-curve/multiverse meta-analysis the Which and How factors constituting the specifications cannot be adopted automatically: rather, they need to be tailor-made each time anew, informed by specific debates in the primary literature or by prior related meta-analyses.

Still, this leaves room for subjectivity (researcher degrees of freedom) and disagreement about what the relevant and reasonable specifications are. Like primary studies, conventional meta-analyses increasingly are preregistered. This could also be done for meta-analytic specification designs. Relatedly, higher consensus might also be achieved by diversifying specification decisions via web-based frameworks, such as community-augmented meta-analysis (Tsuji, Bergmann, & Cristia, 2014) and Curate Science (LeBel, McCarthy, Earp, Elson, & Vanpaemel, 2018). Adversarial collaborations might be expedient (Kahneman, 2003; Kerr, Ao, Hogg, & Zhang, 2018). Combinatorial meta-analysis may act as the final arbiter in such matters.

It has been observed that early decisions in meta-analyses (foremost, the study inclusion-exclusion criteria) frequently generate more result variation than the subsequent statistical modeling (Goodyear-Smith, van Driel, Arroll, & Del Mar, 2012). In other words, the Which factors take precedence over the How factors. Others have noted just the opposite pattern (e.g., Young & Holsteen, 2017; albeit for primary data analyses). It will therefore be interesting to see which importance relations between Which versus How factors will typically arise in applications of specification-curve/multiverse meta-analysis.

We feel confident that there are abundant instances of empirical research suited for, and worthy the effort of, specification-curve/multiverse meta-analysis. We briefly allude here to just three such examples, all taken from current psychological research.

> *Example 1*: Are there ovulatory-cycle effects on women's mating preferences, as predicted by evolutionary psychological theorizing? Yes, according to one meta-analysis, published in the premier journal *Psychological Bulletin* (Gildersleeve, Haselton, & Fales, 2014a), which however prompted an exchange between commentators (Harris, Pashler, & Mickes, 2014; Wood & Carden, 2014) and the authors (Gildersleeve, Haselton, & Fales, 2014b). No, according to another meta-analysis, published almost simultaneously (Wood, Kressel, Joshi, & Louie, 2014), which counterevidence triggered an even more extensive debate (van Anders, 2014; Brown, Cross, Street, & Brand, 2014; Ferguson, 2014; Hyde & Salk, 2014; Jones, 2014; Wood, 2014).
>
> *Example 2*: The research question of possible associations between brain size and cognitive abilities (IQ) has a long and checkered history. According to a widely cited meta-analysis (McDaniel, 2005), these correlations are substantial. According to the Web of Science database, this report currently ranks within the top-20 most-cited articles out of about 1,800 articles published in the journal *Intelligence* since 1980. Based on a substantially larger corpus of primary studies, and accounting for many hitherto unreported effects, other meta-analysts (Pietschnig, Penke, Wicherts, Zeiler, & Voracek, 2015) have found that these associations are noticeably smaller than previously thought and further show a decline in more recent studies (which would be consistent with stronger publication bias in earlier research). Subsequently, others (Gignac & Bates, 2017) applied alternative study eligibility criteria to the same meta-analytic database (i.e., did not retrieve and assemble new data), and in their meta-analysis of merely a subset of the Pietschnig et al. (2015) database, again observed a larger effect. Of note, the specification justified in Gignac and Bates (2017), even if reasonable, remains just one out of many more specifications that are conceivable.
>
> *Example 3*: Over the years, research about aggressive effects of violent video games has become known for controversies surrounding the veracity of this evidence. It appears that multiple (and throughout highly cited) meta-analyses have not resolved the issue to what extent such effects indeed are real (Anderson & Bushman, 2001; Anderson et al., 2010; Greitemeyer & Mügge, 2014) or more likely due to publication bias (Ferguson, 2007a, 2007b, 2015).

As diverse as these examples may appear on the surface, their in-depth commonalities are more important. These include: (1) the conflicting meta-analyses are rooted in controversies already found in the respective literatures which they attempt to synthesize and clarify; (2) even multiple meta-analyses apparently can fail to resolve contentious issues that pervade corresponding primary research; and (3) this sometimes can lead to debates which, likely by more than a few in the research community, are viewed as agonizing and fruitless. We see potential in the proposed approach to mitigate and countersteer against such detrimental phenomena and undesired developments.

In conclusion, whether it be primary studies or meta-analyses, there often seems to be a lack of consensus about *which* data to analyze and *how* to analyze them. Paralleling the potential of specification-curve analysis and multiverse analysis for clarifying the trustworthiness and robustness of evidence from primary studies, an analogously pursued approach to meta-analysis, as introduced here, holds similar promise. Instead of presenting just one meta-analysis and then defending this specification of one's own (or criticizing others' alternative specifications), better assess all possible study subsets (combinatorial meta-analysis) and focus on relevant and justifiable meta-analytic specifications (specification-curve and multiverse meta-analysis).

References

Anderson, C. A., & Bushman, B. J. (2001). Effects of violent video games on aggressive behavior, aggressive cognition, aggressive affect, physiological arousal, and prosocial behavior: A meta-analytic review of the scientific literature. *Psychological Science, 12*, 353–359. https://doi.org/10.1111/1467-9280.00366

Anderson, C. A., Shibuya, A., Ihori, N., Swing, E. L., Bushman, B. J., Sakamoto, A., . . . Saleem, M. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in Eastern and Western countries: A meta-analytic review. *Psychological Bulletin, 136*, 151–173. https://doi.org/10.1037/a0018251

Babková Durdiaková, J., Celec, P., Koborová, I., Sedláčková, T., Minárik, G., & Ostatníková, D. (2017). How do we love? Romantic love style in men is related to lower testosterone levels. *Physiological Research, 66*, 695–703.

Bakkensen, L. A., & Larson, W. D. (2014). Population matters when modeling hurricane fatalities [Letter to the editor]. *Proceedings of the National Academy of Sciences of the United States of America, 111*, E5331–E5332. https://doi.org/10.1073/pnas.1417030111

Berenbaum, S. A., & Beltz, A. M. (2011). Sexual differentiation of human behavior: Effects of prenatal and pubertal organizational hormones. *Frontiers in Neuroendocrinology, 32*, 183–200. https://doi.org/10.1016/j.yfrne.2011.03.001

Breedlove, S. M. (2010). Organizational hypothesis: Instances of the fingerpost. *Endocrinology, 151*, 4116–4122. https://doi.org/10.1210/en.2010-0041

Brown, G. R., Cross, C. P., Street, S. E., & Brand, C. O. (2014). Comment: Beyond "evolutionary versus social": Moving the cycle shift debate forward. *Emotion Review, 6*, 250–251. https://doi.org/10.1177/1754073914523050

Butovskaya, M. L., Vasilyev, V. A., Lazebny, O. E., Burkova, V. N., Kulikov, A. M., Mabulla, A., . . . Ryskov, A. P. (2012). Aggression, digit ratio, and variation in the androgen receptor, serotonin transporter, and dopamine D4 receptor genes in African foragers: The Hadza. *Behavior Genetics, 42*, 647–662. https://doi.org/10.1007/s10519-012-9533-2

Chang, S., Skakkebæk, A., Trolle, C., Bojesen, A., Hertz, J. M., Cohen, A., . . . Gravholt, C. H. (2015). Anthropometry in Klinefelter syndrome: Multifactorial influences due to CAG length, testosterone treatment and possibly intrauterine hypogonadism. *Journal of Clinical Endocrinology and Metabolism, 100*, E508–E517. https://doi.org/10.1210/jc.2014-2834

Cheng, F., Zhao, J., Lu, H., Liu, D., & Liu, L. (2016). The association of the digit ratio and androgen receptor gene CAG polymorphism in patients with premature ovarian failure [in Chinese]. *Journal of Ningxia Medical University, 38*, 856–859, 867.

Christensen, B., & Christensen, S. (2014). Are female hurricanes really deadlier than male hurricanes? [Letter to the editor]. *Proceedings of the National Academy of Sciences of the United States of America, 111*, E3497–E3498. https://doi.org/10.1073/pnas.1410910111

Cohen-Bendahan, C. C. C., van de Beek, C., & Berenbaum, S. A. (2005). Prenatal sex hormone effects on child and adult sex-typed behavior: Methods and findings. *Neuroscience and Biobehavioral Reviews, 29*, 353–384. https://doi.org/10.1016/j.neubiorev.2004.11.004

De Naeyer, H., Bogaert, V., De Spaey, A., Roef, G., Vandewalle, S., Derave, W., . . . Kaufman, J. M. (2014). Genetic variations in the androgen receptor are associated with steroid concentrations and anthropometrics but not with muscle mass in healthy young men. *PLoS One, 9*, e86235. https://doi.org/10.1371/journal.pone.0086235

Durdiaková, J., Celec, P., Laznibatová, J., Minárik, G., Lakatošová, S., Kubranská, A., & Ostatníková, D. (2015). Differences in salivary testosterone, digit ratio and empathy between intellectually gifted and control boys. *Intelligence, 48*, 76–84. https://doi.org/10.1016/j.intell.2014.11.002

Durdiaková, J., Celec, P., Laznibatová, J., Minárik, G., & Ostatníková, D. (2016). Testosterone metabolism: A possible biological underpinning of non-verbal IQ in intellectually gifted girls. *Acta Neurobiologiae Experimentalis, 76*, 66–74. https://doi.org/10.21307/ane-2017-006

Durdiaková, J., Lakatošová, S., Kubranská, A., Laznibatová, J., Ficek, A., Ostatníková, D., & Celec, P. (2013). Mental rotation in intellectually gifted boys is affected by the androgen receptor CAG repeat polymorphism. *Neuropsychologia, 94*, 1693–1698. https://doi.org/10.1016/j.neuropsychologia.2013.05.016

Ferguson, C. J. (2007a). Evidence for publication bias in video game violence effects literature: A meta-analytic review. *Aggression and Violent Behavior, 12*, 470–482. https://doi.org/10.1016/j.avb.2007.01.001

Ferguson, C. J. (2007b). The good, the bad and the ugly: A meta-analytic review of positive and negative effects of violent video games. *Psychiatric Quarterly, 78*, 309–316. https://doi.org/10.1007/s11126-007-9056-9

Ferguson, C. J. (2014). Comment: Why meta-analyses rarely resolve ideological debates. *Emotion Review, 6*, 251–252. https://doi.org/10.1177/1754073914523046

Ferguson, C. J. (2015). Do angry birds make for angry children? A meta-analysis of video game influences on children's and adolescents' aggression, mental health, prosocial behavior, and academic performance. *Perspectives on Psychological Science, 10*, 646–666. https://doi.org/10.1177/1745691615592234

Folland, J. P., McCauley, T. M., Phypers, C., Hanson, B., & Mastana, S. S. (2012). Relationship of 2D:4D finger ratio with muscle strength, testosterone, and androgen receptor CAG repeat genotype. *American Journal of Physical Anthropology, 148*, 81–87. https://doi.org/10.1002/ajpa.22044

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*, 460–465. https://doi.org/10.1511/2014.111.460

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

Gildersleeve, K., Haselton, M. G., & Fales, M. R. (2014a). Do women's mate preferences change across the ovulatory cycle? A meta-analytic review. *Psychological Bulletin, 140*, 1205–1259. https://doi.org/10.1037/a0035438

Gildersleeve, K., Haselton, M. G., & Fales, M. R. (2014b). Meta-analyses and *p*-curves support robust cycle shifts in women's mate preferences: Reply to Wood and Carden (2014) and Harris, Pashler, and Mickes (2014). *Psychological Bulletin, 140*, 1272–1280. https://doi.org/10.1037/a0037714

Goodyear-Smith, F. A., van Driel, M. L., Arroll, B., & Del Mar, C. (2012). Analysis of decisions made in meta-analyses of depression screening and the risk of confirmation bias: A case study. *BMC Medical Research Methodology, 12*, 76. https://doi.org/10.1186/1471-2288-12-76

Greitemeyer, T., & Mügge, D. O. (2014). Video games do affect social outcomes: A meta-analytic review of the effects of violent and prosocial video game play. *Personality and Social Psychology Bulletin, 40*, 578–589. https://doi.org/10.1177/0146167213520459

Habre, C., Tramèr, M. R., Pöpping, D. M., & Elia, N. (2014). Ability of a meta-analysis to prevent redundant research: Systematic review of studies on pain from propofol injection. *British Medical Journal, 349*, g5219. https://doi.org/10.1136/bmj.g5219

Hampson, E., & Sankar, J. S. (2012). Re-examining the Manning hypothesis: Androgen receptor polymorphism and the 2D:4D digit ratio. *Evolution and Human Behavior, 33*, 557–561. https://doi.org/10.1016/j.evolhumbehav.2012.02.003

Harris, C. R., Pashler, H., & Mickes, L. (2014). Elastic analysis procedures: An incurable (but preventable) problem in the fertility effect literature. Comment on Gildersleeve, Haselton, and Fales (2014). *Psychological Bulletin, 140*, 1260–1264. https://doi.org/10.1037/a0036478

Hennig, J., & Rammsayer, T. (2007). Research on 2D:4D: A promising challenge for the study of individual differences [Editorial]. *Journal of Individual Differences, 28*, 53–54. https://doi.org/10.1027/1614-0001.28.2.53

Hines, M. (2010). Sex-related variation in human behavior and the brain. *Trends in Cognitive Sciences, 14*, 448–456. https://doi.org/10.1016/j.tics.2010.07.005

Hines, M. (2011). Gender development and the human brain. *Annual Review of Neuroscience, 34*, 69–88. https://doi.org/10.1146/annurev-neuro-061010-113654

Hofmann, B. (2018). Fake facts and alternative truths in medical research. *BMC Medical Ethics, 19*, 4. https://doi.org/10.1186/s12910-018-0243-z

Hönekopp, J. (2013). No evidence that 2D:4D is related to the number of CAG repeats in the androgen receptor gene. *Frontiers in Endocrinology, 4*, 185. https://doi.org/10.3389/fendo.2013.00185

Hönekopp, J., & Watson, S. (2010). Meta-analysis of digit ratio 2D:4D shows greater sex difference in the right hand. *American Journal of Human Biology, 22*, 619–630.

Hurd, P. L., Vaillancourt, K. L., & Dinsdale, N. L. (2011). Aggression, digit ratio and variation in androgen receptor and monoamine oxidase A genes in men. *Behavior Genetics, 41*, 543–556. https://doi.org/10.1007/s10519-010-9404-7

Hyde, J. S., & Salk, R. H. (2014). Comment: Menstrual cycle fluctuations in women's mate preferences. *Emotion Review, 6*, 253–254. https://doi.org/10.1177/1754073914523049

Ioannidis, J. P. (2016). The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *Milbank Quarterly, 94*, 485–514. https://doi.org/10.1111/1468-0009.12210

Jones, B. C. (2014). Comment: Alternatives to Wood et al.'s conclusions. *Emotion Review, 6*, 254–256. https://doi.org/10.1177/1754073914523048

Jung, K., Shavitt, S., Viswanathan, M., & Hilbe, J. M. (2014a). Female hurricanes are deadlier than male hurricanes. *Proceedings of the National Academy of Sciences of the United States of America, 111*, 8782–8787. https://doi.org/10.1073/pnas.1402786111

Jung, K., Shavitt, S., Viswanathan, M., & Hilbe, J. M. (2014b). Reply to Bakkensen and Larson: Population may matter but does not alter conclusions [Letter to the editor]. *Proceedings of the National Academy of Sciences of the United States of America, 111*, E5333. https://doi.org/10.1073/pnas.1419330111

Jung, K., Shavitt, S., Viswanathan, M., & Hilbe, J. M. (2014c). Reply to Christensen and Christensen and to Malter: Pitfalls of erroneous analyses of hurricanes names [Letter to the editor]. *Proceedings of the National Academy of Sciences of the United States of America, 111*, E3499–E3500. https://doi.org/10.1073/pnas.1411652111

Jung, K., Shavitt, S., Viswanathan, M., & Hilbe, J. M. (2014d). Reply to Maley: Yes, appropriate modeling of fatality counts confirms female hurricanes are deadlier [Letter to the editor]. *Proceedings of the National Academy of Sciences of the United States of America, 111*, E3835. https://doi.org/10.1073/pnas.1414111111

Kahneman, D. (2003). Experiences of collaborative research. *American Psychologist, 58*, 723–730. https://doi.org/10.1037/0003-066X.58.9.723

Kerr, N. L., Ao, X., Hogg, M. A., & Zhang, J. (2018). Addressing replicability concerns via adversarial collaboration: Discovering hidden moderators of the minimal intergroup discrimination effect. *Journal of Experimental Social Psychology, 78*, 66–76. https://doi.org/10.1016/j.jesp.2018.05.001

Knickmeyer, R. C., Woolson, S., Hamer, R. M., Konneker, T., & Gilmore, J. H. (2011). 2D:4D ratios in the first 2 years of life: Stability and relation to testosterone exposure and sensitivity. *Hormones and Behavior, 60*, 256–263. https://doi.org/10.1016/j.yhbeh.2011.05.009

Kubranská, A., Lakatošová, S., Schmidtová, E., Durdiaková, J., Celec, P., & Ostatníková, D. (2014). Spatial abilities are not related to testosterone levels and variation in the androgen receptor in healthy young men. *General Physiology and Biophysics, 33*, 311–319. https://doi.org/10.4149/gpb_2014005

Kühberger, A., Scherndl, T., Ludwig, B., & Simon, D. M. (2016). Comparative evaluation of narrative reviews and meta-analyses: A case study. *Zeitschrift für Psychologie, 224*, 145–156. https://doi.org/10.1027/2151-2604/a000250

Latourelle, S. M., Elwess, N. L., & Elwess, J. M. (2008). Finger forecasting: A pointer to athletic prowess in women – a preliminary investigation by an undergraduate biology class. *American Biology Teacher, 70*, 411–414. https://doi.org/10.1662/0002-7685(2008)70[411:FFAPTA]2.0.CO;2

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science, 1*, 389–402. https://doi.org/10.1177/2515245918787489

Loehlin, J. C., Medland, S. E., & Martin, N. G. (2012). Is CAG sequence length in the androgen receptor gene correlated with finger-length ratio? *Personality and Individual Differences, 52*, 224–227. https://doi.org/10.1016/j.paid.2011.09.009

Maley, S. (2014). Statistics show no evidence of gender bias in the public's hurricane preparedness [Letter to the editor]. *Proceedings of the National Academy of Sciences of the United States of America, 111*, E3834. https://doi.org/10.1073/pnas.1413079111

Malter, D. (2014). Female hurricanes are not deadlier than male hurricanes [Letter to the editor]. *Proceedings of the National Academy of Sciences of the United States of America, 111*, E3496. https://doi.org/10.1073/pnas.1411428111

Manning, J. T., Bundred, P. E., Newton, D. J., & Flanagan, B. F. (2003). The second to fourth digit ratio and variation in the androgen receptor gene. *Evolution and Human Behavior, 24*, 399–405. https://doi.org/10.1016/S1090-5138(03)00052-7

Manning, J. T., Scutt, D., Wilson, J., & Lewis-Jones, D. I. (1998). The ratio of 2nd to 4th digit length: A predictor of sperm numbers and concentrations of testosterone, luteinizing hormone and oestrogen. *Human Reproduction, 13*, 3000–3004. https://doi.org/10.1093/humrep/13.11.3000

Mas, M., Alonso, C., Hernandez, P., Fernandez, M., Gutierrez, P., Salido, E., & Baez, D. (2009). Androgen receptor CAG and GGN polymorphisms and 2D:4D finger ratio in male to female transsexuals [Abstract]. *Journal of Sexual Medicine, 6*(Suppl. 5), 419–420.

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

Naudet, F., Schuit, E., & Ioannidis, J. P. A. (2017). Overlapping network meta-analyses on the same topic: Survey of published studies. *International Journal of Epidemiology, 46*, 1999–2008. https://doi.org/10.1093/ije/dyx138

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69*, 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

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

Patel, C. J., Burford, B., & Ioannidis, J. P. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology, 68*, 1046–1058. https://doi.org/10.1016/j.jclinepi.2015.05.029

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

Rohrer, J. M., Egloff, B., & Schmukle, S. C. (2017). Probing birth-order effects on narrow traits using specification-curve analysis. *Psychological Science, 28*, 1821–1832. https://doi.org/10.1177/0956797617723726

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

Silberzahn, R., Uhlmann, E. L., Martin, D., Anselmi, P., Aust, F., Awtrey, E. C., . . . Nosek, B. A. (2018). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. *Advances in Methods and Practices in Psychological Science, 1*, 337–356. https://doi.org/10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366. https://doi.org/10.1177/0956797611417632

Simmons, J., Nelson, L., & Simonsohn, U. (2012). A 21 word solution. *Dialogue, 26*, 4–7. https://doi.org/10.2139/ssrn.2160588

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). *Specification curve: Descriptive and inferential statistics on all reasonable specifications*. Retrieved from http://sticerd.lse.ac.uk/seminarpapers/psyc16022016.pdf

Smith, G. (2016). Hurricane names: A bunch of hot air? *Weather and Climate Extremes, 12*, 80–84. https://doi.org/10.1016/j.wace.2015.11.006

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*, 702–712. https://doi.org/10.1177/1745691616658637

Taylor, A. E., & Munafò, M. R. (2016). Triangulating meta-analyses: The example of the serotonin transporter gene, stressful life events and major depression. *BMC Psychology, 4*, 23. https://doi.org/10.1186/s40359-016-0129-0

Tsuji, S., Bergmann, C., & Cristia, A. (2014). Community-augmented meta-analyses: Toward cumulative data assessment. *Perspectives on Psychological Science, 9*, 661–665. https://doi.org/10.1177/1745691614552498

van Anders, S. M. (2014). Comment: The social neuroendocrinology example: Incorporating culture resolves biobehavioral evolutionary paradoxes. *Emotion Review, 6*, 256–257. https://doi.org/10.1177/1754073914523047

Voracek, M. (2011). Special issue preamble: Digit ratio (2D:4D) and individual differences research. *Personality and Individual Differences, 51*, 367–370. https://doi.org/10.1016/j.paid.2011.04.018

Voracek, M. (2014). No effects of androgen receptor gene CAG and GGC repeat polymorphisms on digit ratio (2D:4D): A comprehensive meta-analysis and critical evaluation of research. *Evolution and Human Behavior, 35*, 430–437. https://doi.org/10.1016/j.evolhumbehav.2014.05.009

Voracek, M., Kaden, A., Kossmeier, M., Pietschnig, J., & Tran, U. S. (2018). Meta-analysis shows associations of digit ratio (2D:4D) and transgender identity are small at best. *Endocrine Practice, 24*, 386–390. https://doi.org/10.4158/EP-2017-0024

Voracek, M., & Loibl, L. M. (2009). Scientometric analysis and bibliography of digit ratio (2D:4D) research, 1998–2008. *Psychological Reports, 104*, 922–956. https://doi.org/10.2466/PR0.104.3.922-956

Voracek, M., Pietschnig, J., Nader, I. W., & Stieger, S. (2011). Digit ratio (2D:4D) and sex-role orientation: Further evidence and meta-analysis. *Personality and Individual Differences, 51*, 417–422. https://doi.org/10.1016/j.paid.2010.06.009

Voracek, M., Tran, U. S., & Dressler, S. G. (2010). Digit ratio (2D:4D) and sensation seeking: New data and meta-analysis. *Personality and Individual Differences, 48*, 72–77. https://doi.org/10.1016/j.paid.2009.08.019

Warrington, N. M., Shevroja, E., Hemani, G., Hysi, P. G., Jiang, Y., Auton, A., . . . Evans, D. M. (2018). Genome-wide association study identifies nine novel loci for 2D:4D finger ratio, a putative retrospective biomarker of testosterone exposure *in utero*. *Human Molecular Genetics, 27*, 2025–2038. https://doi.org/10.1093/hmg/ddy121

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., & Bouwman, J. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*, 160018. https://doi.org/10.1038/sdata.2016.18

Wood, W. (2014). Author reply: Once again, menstrual cycles and mate preferences. *Emotion Review, 6*, 258–260. https://doi.org/10.1177/1754073914523053

Wood, W., & Carden, L. (2014). Elusiveness of menstrual cycle effects on mate preferences: Comment on Gildersleeve, Haselton, and Fales (2014). *Psychological Bulletin, 140*, 1265–1271. https://doi.org/10.1037/a0036722

Wood, W., Kressel, L., Joshi, P. D., & Louie, B. (2014). Meta-analysis of menstrual cycle effects on women's mate preferences. *Emotion Review, 6*, 229–249. https://doi.org/10.1177/1754073914523073

Young, C. (2018). Model uncertainty and the crisis in science. *Socius*. Advance online publication. https://doi.org/10.1177/2378023117737206

Young, C., & Holsteen, K. (2017). Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods & Research, 46*, 3–40. https://doi.org/10.1177/0049124115610347

Zhang, C., Dang, J., Pei, L., Guo, M., Zhu, H., Qu, L., . . . Huo, Z. (2013). Relationship of 2D:4D finger ratio with androgen receptor CAG and GGN repeat polymorphism. *American Journal of Human Biology, 25*, 101–106. https://doi.org/10.1002/ajhb.22347

Zhang, C., Lu, H., Hao, S., Yan, Y., Dang, J., Zheng, L., . . . Huo, Z. (2016). Relationship between androgen receptor CAG/GGN repeat polymorphisms and the ratio of 2D:4D [in Chinese]. *Acta Anatomica Sinica, 47*, 409–414.

Zhang, K., Yang, X., Yang, Y., Xue, M., Fang, P., Wang, B., . . . Gong, P. (2018). *Revisiting the relation of ratio of 2D:4D with the androgen receptor (AR) gene and the circulating testosterone levels: Cross-sectional study and meta-analyses*,. Manuscript submitted for publication

Zhang, Z., Kossmeier, M., Tran, U. S., Voracek, M., & Zhang, H. (2017). Rainforest plots for the presentation of patient-subgroup analysis in clinical trials. *Annals of Translational Medicine, 5*, 24. https://doi.org/10.21037/atm.2017.10.07

History
Received February 28, 2018
Revision received October 27, 2018
Accepted October 28, 2018
Published online March 29, 2019

Acknowledgments

The authors gratefully acknowledge the input generated by inspiring audience questions and discussions at the Research Synthesis 2018 Conference (June 2018, Leibniz Institute for Psychology Information, University of Trier, Germany) and at the COSB (Centre for Organismal Systems Biology) Colloquium (May 2018, University of Vienna, Austria), where earlier versions of this work were presented, as well as excellent anonymous reviewer comments, all of which helped to improve the presentation of this work.

Martin Voracek
Department of Basic Psychological Research and Research Methods
Faculty of Psychology
University of Vienna
Liebiggasse 5
1010 Vienna
Austria
martin.voracek@univie.ac.at

Michael Kossmeier
Department of Basic Psychological Research and Research Methods
Faculty of Psychology
University of Vienna
Liebiggasse 5
1010 Vienna
Austria
michael.kossmeier@univie.ac.at