MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

Occurrence and awareness of decline effects and publication bias in clinical child psychology: Causes and consequences.

verfasst von / submitted by

Junia Eder, BSc

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

Master of Science (MSc)

Wien, 2023 / Vienna 2023

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

Table of Contents

Acknowledgements

I'd like to say a big thank you to everybody who helped and assisted me on the way.

Firstly, a very big tank you to my supervisor Jakob Pietschnig. Thank you so much for all your support, your patience with me and all the amazing opportunities that you have offered me in the past. Thank you for listening to my sorrows when I had trouble and thank you for not giving up on me.

Heartfelt thanks also to Magdalena Siegel, my former work colleague, who taught me basically everything I know about meta-analyses. Thank you for looking after me and your constant and kind advice and support. I really appreciate it!

Thank you, Stefan Keinrath, my master's buddy. Thank you for all the countless days of coding, discussing, bug fixing and advice and for being my second coder for my studies. It has been a pleasure to do this with you together.

Thank you, Julia Egger, for proofreading my manuscript and all your insightful comments.

I'd also like to thank Steven Pinker, for giving me the pep talk I needed.

Thank you, John Protzko for being an awesome guy! It's been a pleasure to have met you and thank you for your willingness to collaborate with me and help me with my little survey.

I'd also like to express my gratitude to all the researchers who took interest in my survey. Thank you for your time and your answers, I know you are all really busy people, so I appreciate it even more!

Finally, the biggest thanks go out to my family. Thank you, mum and dad for being the constant loving background humming of my life. Thank you for being unconditionally supporting of whatever crazy idea enters my mind, for countless advice and understanding. I love you both.

The last thank you belongs to Valentin. Thank you for helping me to debug code again and again, to think statistical theorems through with me you are only vaguely aware of, for all your smart remarks and all your love and calmness and outside support that kept me sane. Thank you for being you. I love you.

Abstract (English)

Since the replication crisis hit science, we have observed the phenomenon that in various scientific fields our effects seem to diminish. This has become known as decline effect. One of the main postulated drivers for decline effect is publication bias. While decline effects and publication bias have been found in psychology and its subdisciplines, nobody has assessed the field of clinical child psychology so far. For this study, we meta-meta-analysed over 40 meta-analyses (k = 1,000+, N = 1,000,000+) that were published in the "Journal of Child Psychology and Psychiatry". Effect misestimations in the form of declining and increasing effects were found as well as evidence for publication bias. Various study characteristics seemed to influence bias but indices regarding the publication performance of a study (e.g., number of citations) did not. Publication bias but also various genuine clinical reasons (e.g., changes in diagnostic criteria) seemed to drive the effect misestimation. A survey showed that clinical child researchers were aware of publication bias and decline effect – but the publication bias detection methods used were often outdated. In the future, various steps should be undertaken to improve scientific rigor (e.g., open data, open syntax, pre-registration) and make it more appealing to researchers to publish their findings regardless of the outcome.

Abstract (Deutsch)

Seit die Replikationskrise die Wissenschaft traf, haben wir das Phänomen beobachtet, dass unsere Effekte immer kleiner werden. Das wurde bekannt unter dem Namen *decline effect*. Eine der postulierten Hauptgründe für decline effect scheint *publication bias* zu sein. Auch in der Psychologie und ihren Subdisziplinen hat man decline effect und publication bias bereits gefunden. Bisher hat niemand das Feld der klinischen Kinderpsychologie hinsichtlich dessen untersucht. Für diese Studie haben wir 40 Meta Analysen (k = 1,000+, N = 1,000,000+) die im „Journal of Child Psychology and Psychiatry“ publiziert wurden, meta-meta analysiert. Effektverzerrungen in Form von Abnahmen und Zunahmen wurden gefunden, sowie Evidenz für publication bias. Verschiedene Studiencharakteristia scheinen einen Einfluss auf die Effektverzerrung zu haben, Indizes, die den Publikationserfolg einer Studie anzeigen (z.B. Anzahl der Zitationen) zeigten keinen. Publication bias, aber auch verschiedene genuine klinische Gründe (z.B. Änderungen bei Diagnosekriterien), scheinen die Effektverzerrung zu verursachen. Eine Umfrage unter Wissenschaftler*innen mit klinisch-kinderpsycholgischem Hintergrund hat gezeigt, dass diese von publication bias und decline effect grundsätzlich Bescheid wussten – aber die Detektionsmethoden für publication bias, die verwendet wurden, waren oft veraltet. Für die Zukunft sollten verschiedene Schritte unternommen werden, um wissenschaftliche Studien zu verbessern (z.B. Open Data, Open Syntax, pre-registrieren) und es ansprechender für Wissenschaftler*innen zu machen, ihre Erkenntnisse unabhängig ihrer Ergebnisse zu publizieren.

1. Introduction

Science comes from innovation, meaning that new ideas and exciting findings act as drive for research. But one must not forget that almost as important as discovering new findings is to establish robust evidence for already existing findings. Findings can also be refuted later if the so-claimed effect cannot be found in future studies (Ioannidis, 2005).

A good method to build up scientific evidence is by replicating a study. This allows us to confirm whether we can find the same effect with the same methodology in a slightly different setting. If we are dealing with a "true" effect, we should be able to consistently find it in several studies.

But over the last decades, researchers have found that previously discovered effects do not replicate anymore or at least not in the same magnitude. This has led to the so-called *replication crisis* (Protzko & Schooler, 2017) – previously established effects fail to replicate in subsequent studies and the effects found later are either smaller or even non-existent (Fanelli et al., 2017; Schooler, 2011). This became known as *decline effect* (Protzko & Schooler, 2017).
It is also worth mentioning that effects do not always decline. It is also possible that effects can increase over time (*increase effect*) or just remain stable (*stable effects*).

A special form of decline effect is called the *Proteus effect*. It occurs when the effect changes its sign over time (Pietschnig et al., 2019; Ioannidis & Trikolinos, 2005). This can be seen as an extreme case of decline effect, as the initial effect was not only inflated but even pointing in the wrong direction.

1.1 Prevalence of decline effect

Ioannidis (2005) estimated that the majority of our scientific findings today are false or inflated. And indeed, decline effect can be found in almost all scientific fields e.g., biology (Clements et al., 2022; Jeschke et al., 2012), medicine (van Aert et al., 2019; Begg & Berlin, 1988), genetics (Siontis et al., 2010), psychology (Bakker et al., 2012), but also in computer science (Cockburn et al., 2020), and economics (Ioannidis et al., 2017).

The Open Science Collaboration (2015) published a paper estimating the reproducibility of psychological science. Replications of 100 experimental and correlational studies published in three top-tier psychology journals were conducted. They found that the

mean effect size of the replicated effects (Mean r = .197, SD = .257) was only half the magnitude than those of the mean original effects (Mean r = .403, SD = .188), thereby demonstrating an impressive decline effect. Ninety-seven percent of the original studies had statistically significant results, whereas only 36% of the replicated ones were statistically significant.

Certain subdisciplines in psychology have already been investigated regarding decline effect and non-replicability. Examples include intelligence research (Nuijten et al., 2020; Pietschnig et al., 2019; Gong & Jiao, 2019), industrial/organizational psychology (Siegel et al., 2022; Banks & McDaniel, 2011), or experimental psychology (Francis, 2012).

When it comes to the field of clinical psychology, there is evidence hinting that decline effect may be present here as well. For example, declining effects were already found in general clinical research (Easterbrook et al., 1991; Begg & Berlin, 1989), or in treating depression with psychotherapy (Cuijpers et al., 2010).
Clinical psychology may be especially vulnerable to declining and initially inflated results, because of the nature of clinical studies itself. Clinical studies usually investigate very specific populations (e.g., people with a certain disorder) in a very specific setting (e.g., a hospital ward). Experiments usually have quite simple designs, meaning that often there are few case numbers and small sample sizes (Ioannidis & Trikolinos, 2005) with an average of about 15-20 participants (Blease et al., 2023). This leads to low study power and potentially biased results and declining effects. Another aspect in clinical experiments is the distinction between the "(patient) experimental group" and the "(healthy) control group". The groups are distinguished by using diagnostic criteria, usually according to the *International Classification of Diseases* (*ICD-11*, World Health Organization, 2022) or the *Diagnostic and Statistical Manual of Mental Disorder* (*DSM-5*, American Psychiatric Association, 2013). As these diagnostic instruments change over time, this variation could impact clinical empirical research, as people may be getting a slightly different diagnosis today than they would have 20 years ago.

Bias in clinical psychology is especially problematic as one of the best-practice recommendations today is *evidence-based practice* – meaning that treatments should orient themselves on evidence-based results (DePalma, 2000). If these results were biased, this would have grave consequences for the patients as they might be given the wrong treatment based on erroneous evidence (Goldacre et al., 2016). In medicine, this phenomenon is called

medical reversals - medical practices that were considered useful but later were discovered to be ineffective or harmful e.g., the case of "aprotinin" (Blease et al., 2023).

When it comes to clinical *child* psychology, almost no studies can be found that investigate possible bias or decline effect for this field. However, we have tentative evidence hinting that bias may be present here as well, e.g., McLeod and Weisz (2004) found that the overall effect size in youth psychotherapy dissertations is less than half than those in published studies. But apart from that, studies are scarce.

1.2 Reasons for decline effect

Several reasons for decline effect have been discussed in the literature. We can distinguish between genuine reasons and non-genuine reasons for decline effect.

1.2.1 Genuine declines

Sometimes temporal effect changes can occur because of genuine societal and global changes and developments. A prominent example is the *Flynn effect* (Flynn, 1987). The Flynn effect describes the phenomenon of the worldwide rise of IQ over the last decades, with an average global increase of 3 IQ points per decade (Pietschnig & Voracek, 2015). Research has shown that the reasons for the Flynn effect seem to be caused by environmental factors, such as improved worldwide nutrition (Lynn, 2009), more global access to education and better schooling (Baker et al., 2015), as well as more habituation and practice with cognitive ability tests, such as PISA or assessment centers in job recruitment (Neisser, 1997).
Newest findings show that the Flynn effect has been stagnating in the last few years and even reversing itself (Pietschnig & Gittler, 2015). It seems that nutrition, education, and other environmental factors can only push our global intelligence so far, and that a ceiling effect can be observed.
It is important to point out that genuine declines, as observed with the Flynn effect, usually refer to temporal changes of means and not effect sizes. Means are usually not affected by mechanisms of non-genuine declines, as their publication performance does not rely on a significant outcome (e.g., mean temporal changes in global intelligence are publishable whether intelligence is increasing or decreasing).

1.2.2 Non-genuine declines

For this study, we will focus primarily on the non-genuine reasons for temporal effect changes. Apart from genuine effect changes, methodological issues can also have an influence on our scientific findings. Some of them are outlined in the following sub-sections.

1.2.2.1 Underpowered studies

When studies are conducted with low study power the possibility for statistically significant large effects increases. But these effects are not driven by a true effect but from increased error variance (which rises with lower study power), and thereby increasing the possibility for statistical "flukes" (Button et al., 2013; Protzko & Schooler, 2017). Subsequent studies usually use larger sample sizes and therefore report smaller effect sizes (Ioannidis, 2005). Concerning the field of psychology, most empirical psychological studies are chronically underpowered (with median sample sizes from 40 – 48 for group comparison tests; Wetzels et al., 2011) and within the clinical field even more (on average 15 – 20 participants per investigation; Blease et al., 2023). Researchers in psychology tend to adhere to, among other things, previous research findings or the average sample size that is common for the field when assessing power. Considering that the initial effect might be an inflated statistical fluke, subsequent research only produces further low-powered findings, when power is not estimated properly (Vankov et al., 2014).

1.2.2.2 Regression to the mean

Some scientists argue that decline effect is an attribute of the phenomenon we know as "regression to the mean". After an exaggerated, possibly inflated, finding, smaller effects follow so that on average the effect sizes will regress to the mean, thereby declining from the initial effect (Schooler, 2011).

1.2.2.3 Changes in methodology

As scientific methodology develops, we are able to investigate our topics of interest more precisely. More fine-grained methods allow us to delve deeper into the topic and assess them in a more differentiated manner. When a topic is investigated over time with better methods, a decline effect can occur because these methods produce more accurate and therefore often smaller effects (Protzko & Schooler, 2017).

1.2.2.4 Hotness of field

Ioannidis (2005) has shown that the more exciting a new scientific finding is, the greater the probability that these findings are false. Hotness of field means that a certain topic or trend sparks a considerable amount of interest, driving scientific findings in a certain direction. Since everybody is "hooked" by a certain theory, it will become harder to publish studies with contradicting findings. Confirmation bias may also play a role here as findings become "canonised" over time (Blease et al., 2023). This might encourage selective publishing when contradictory findings and failed replications are rejected for publication ("file drawer effect", Rosenthal, 1979). When conflicting findings start to accumulate to a certain level that it cannot be ignored anymore, canonisation usually comes to an end.

1.2.2.5 Publication bias

Publication bias means that findings published in scientific journals differ systematically from unpublished results. This occurs when mainly statistically significant, hypothesis-conforming studies are published, while studies with contradicting findings or those that support the null-hypothesis are not. Significant and novel findings usually stem from smaller, underpowered studies. Subsequent replications generally have higher statistical demands, i.e., better methodology or higher power and larger sample sizes. Higher power also reduces error variance, therefore, the probability for inflated effect sizes decreases and subsequently, higher-powered studies produce smaller effect sizes.
However, academia promotes selective reporting and publishing of novel and significant findings, as they have a 40% higher rate of getting published than nonsignificant findings (Franco et al., 2014) and are on average published one to two years earlier than nonsignificant ones (Protzko & Schooler, 2017). Moreover, the pressure to publish in academia is high, especially if future career options and/or funding are linked to one's publication performance (Fanelli et al., 2017).

1.2.2.6 Questionable research practices (QRPs)

The publication pressure described above may lead some people to employ certain techniques to make their data more attractive for publication: This includes fiddling with non-significant data long enough that its significance level moves just below the acceptable *p*-value ("*p*-hacking"), conducting numerous experiments but only writing up and publishing the significant ones ("selective reporting" / "publication bias", also called the "file drawer

effect"; Rosenthal, 1979), hypothesising after results are known ("HARKing") or in the worst case even fabricating or inventing data (Neuroskeptic, 2012).

The main aim of QRPs is usually to push a dataset just below the significance level ("*p*-hacking"). But if the *p*-value is artificially reduced, the effect sizes become inflated. Subsequent studies that did not artificially reduce their *p* will have less inflated and, therefore, smaller effect sizes.

1.3 The purpose of this study

Protzko and Schooler (2017, p.101) have stated that:

> Clearly, understanding the implications and magnitude of decline effects requires more field-wide analyses to determine the degree to which decline effects represent a disproportionately large tendency of scientific results over time.

To our knowledge, no one has systematically investigated the field of clinical child psychology regarding possible biases and declining effect sizes. To close this knowledge gap, we conducted a meta-meta-analysis investigating multiple aspects of decline effect and the possible influence of publication bias.

2. Methods

The study was preregistered on OSF: https://osf.io/s5n3m. The supplemental material can be found at: https://osf.io/9c8yf/.

2.1 Research question

On the Meta level (calculations over the studies that are included in a meta-analysis) the main goals were twofold: First, to investigate whether decline effect can be found in the scientific literature of clinical child psychology, and if yes, estimate its size, and possible direction (decline effect or increase effect) and calculate the respective indices for the Meta-Meta level (crude differences and regression slopes) . Second, to examine the occurrence of publication bias and its possible influence on decline effects.

On the Meta-Meta level (the calculations over all meta-analyses), we sought to investigate whether including grey literature (everything that is not a published journal article or a book chapter) impacts decline effect (Schooler, 2011). When studies with small *N* are

included in a meta-analysis, the between-study heterogeneity increases (IntHout et al., 2015). To account for that, associations between heterogeneity estimates (I^2 and tau^2), sample numbers and participant numbers were assessed. Furthermore, the average power of all included studies was calculated, as psychological studies often suffer from low power and this can lead to inflated effect sizes (Button et al., 2013; Protzko & Schooler, 2017). Finally, associations between certain study characteristics and publishing-related characteristics (see Hypotheses) and measures of effect misestimation were investigated on a meta-analytic level.

Furthermore, we conducted an exploratory examination regarding the awareness of temporal effect changes in the scientific field by surveying the relevant research community (IRB Code: #202205).

2.2 Hypotheses

The following hypotheses were investigated:

1.) Crude differences will show more effect decreases than increases.

2.) The average annual change of effect sizes will show a decline.

3.) There will be a difference between decline effects, depending on the nature of the effect size metric (correlational effect sizes vs. effect sizes based on means).

4.) Publication bias methods will detect bias more, if meta-analyses with only published studies are compared to meta-analyses that include grey literature as well.

5a.) I squared will have a *positive association* with sample numbers (*k*), total participant numbers (*N*) and average participant number per study.

5b.) Tau squared will have *negative association* with sample numbers (*k*), total participant numbers (*N*) and average participant number per study.

6a.) Crude differences will have certain associations with the following initial study characteristics:

- absolute effect size (pos.)
- sample size (neg.)
- current Impact Factor initial journal (pos.)
- annual citations per year (pos.)
- publication year (no direction)

6b.) Crude differences will have certain associations with the following meta-analytic characteristics:

- absolute summary effect (no dir.)
- sample number (*k*) (pos.)
- covered time span (no dir.)
- publication year (no dir.)
- type of included study (pub/unpub) (pub only: pos.)

7a.) Regression slopes will have certain associations with the following initial study characteristics:

- absolute effect size (neg.)
- sample size (pos.)
- current Impact Factor initial journal (neg.)
- annual citations per year (neg.)
- publication year (no direction)

7b.) Regression slopes will have certain associations with the following meta-analytic characteristics:

- absolute summary effect (no dir.)
- sample number (*k*) (no dir.)
- covered time span (no dir.)
- publication year (no dir.)
- type of included study (pub/unpub) (pub only: pos.)

2.3 Search Strategy

All meta-analyses published in the *Journal of Child Psychology and Psychiatry* (JCPP) were investigated. The JCPP is one of the top tier journals on all topics concerning clinical child psychology and psychiatry. The journal has been founded with the intention to "bring together original papers concerned with the child, from such diverse disciplines as psychiatry, psychology, paediatrics, psychoanalysis, social case-work and sociology" (Berger & Hersov, 2009, p.2). With an Impact Factor of 8.265 in 2021, it has been on rank 8/80 on the SCIE (Science Citation Index Expanded) and rank 3/78 in the SSCI (Social Science Citation Index) in the Category of “Psychology” (2021 Journal Impact Factor).

While high-ranking journals usually have a good reputation, problems arise when the journal policies favour new and exciting findings. An example is a statement given by the *Journal of Personality and Social Psychology* (Impact Factor = 7.672 in 2020). In 2011, they stated that "[t]his journal does not publish replication studies, whether successful or unsuccessful" (Smith, 2011, cited in Aldhous, 2011) regarding unsuccessful replications of the Bem precognition paper (Bem, 2011). If high-ranking journals strive to only publish new, "hot", and exciting findings and no replications, this might pollute the scientific field with false and inflated effect sizes.

2.3.1 Literature Search

An electronic literature search was conducted to locate all eligible papers. The search string: "(TS=("meta-analy*" OR "research synthes*" OR "systematic review")) AND SO=(JOURNAL OF CHILD PSYCHOLOGY "AND" PSYCHIATRY)" was used to extract all meta-analyses from the *Journal of Child Psychology and Psychiatry*. The first search was conducted on the 28.12.21 and provided 145 results. Full-texts from all studies were obtained and screened for eligibility. If a study fitted the inclusion criteria, it was coded according to a pre-defined standardised coding sheet by the author (JE) and once by a second coder (SK) to ensure reliability of coding. After all studies have been coded once, the literature search was updated once more on 21.06.22 by using the exact same search string which yielded six additional results.
After coding was finished, coding sheets were compared, and inconsistencies discussed between the two coders until consensus was reached. When no consensus could be reached, a third independent researcher (JP) was consulted until agreement was achieved.

All primary data were extracted from the studies for further analyses. If substantial data was missing that permitted reanalyses of the paper, the study authors were contacted, and the missing data was requested via email. After an initial mail, a reminder was sent after two weeks. If there was no answer after that, the study was excluded.

A list with all the included and excluded studies can be found in the supplemental material S1.

2.4 Eligibility Criteria

To be included into this study a meta-analysis had to fit certain inclusion criteria: First, it had to be a standard meta-analysis using classical meta-analytical techniques (e.g., no

network meta-analyses, structural equation modelling, method of correlated vectors, path-coefficients, …). Second, the meta-analyses had to be based on summarising individual different studies coming from different samples and not just be a reanalysis of already existing data. Third, a meta-analysis had to report at least one standardised effect size metric (i.e., Cohens *d*, Hedges *g*, Pearson *r*, Fishers *z*, Odds Ratio, logOdds) to allow recalculation of data. Measures such as Risk Ratios or Hazard Ratios were excluded. Fourth, meta-analyses investigating time trends of means were excluded (cross-temporal meta-analyses), since changes in sample means should not be related to the mechanisms that cause decline effect. Fifth, a meta-analysis had to cover the whole timespan, therefore all meta-analyses that used a time span restriction (e.g., only investigating studies in a certain time span without a conceptual reason) were excluded.

If a paper contained multiple meta-analyses, the one which included the earliest study (i.e., the oldest effect, the so-called "initial study") was chosen. If the earliest study was present in more than one meta-analysis, the second rule was to choose the meta-analysis with the largest *k* (number of effect sizes). If more than one meta-analysis contained the same number of effect sizes, one meta-analysis was chosen at random.

2.5 Coded Variables

Meta-Analysis

For each eligible meta-analysis the following variables were coded: (i) number of meta-analyses in the study, (ii) main research question, (iii) dependent data (yes/no), (iv) model type (fixed/ random), (v) type of meta analyses (Hedges & Olkin/ Hunter & Schmidt)[1], (vi) meta-analytic summary effect and *p*-value (if given), (vii) type of effect size, (viii) I squared/ tau squared, (ix) sample type (healthy/patients/mixed), (x) overall *N*, (xi) number of effect sizes (*k*), (xii) publication bias assessed (yes/no), (xiii) if yes, which methods, (xiv) publication bias outcome (yes/no/some/not investigated), (xv) grey literature inclusion (yes/no), (xvi) covered time span.

[1] Hedges & Olkin's approach will be preferred. In Hunter & Schmidt meta-analyses the uncorrected value (= the observed value) of the meta-analytic summary effect will be reported. If the observed value is not given, the value with the least number of corrections will be taken (e.g., corrections for unreliability, sampling error or range restriction).

Initial Study

For the initial study the following variables were coded: (i) initial effect size and *p*-value (if given), (ii) initial *n*, (iii) citations from the initial journal in "Web of science", (iv) current Impact Factor from the initial journal, (v) does the research question from the meta-analysis match the research question of the initial study (match/ no match), (vi) does the sign from the meta-analysis match the sign of the initial study sign match: meta-analysis/initial study (match/ no match = Proteus effect).

Primary Data

For all eligible meta-analyses that provided the primary data (either by table, supplement in the study, or acquired through author contact), the following variables were extracted for each data point: (i) name of the study, (ii) year published, (iii) effect size, (iv) n, (v) publication status (1 = published, 0 = unpublished)[2].

If specific variable values were missing, they were omitted from further calculations. No data imputation took place.

2.6 Preliminary analyses

Prior to any calculations, all effect sizes (meta-analytic summary effect, initial study effect and the primary study effects) were transformed into Fishers *z*, to obtain a common metric. For further analyses two indices were calculated: crude differences and regression slopes.

2.6.1 Crude differences

Crude differences were calculated by subtracting the meta-analytic summary effect from the initial study effect. This was done for all included studies. In cases where the summary effect and the initial effect of a study are negative, those values were multiplied by (-1), so that *positive values* show a *decline*, whereas *negative values* show an *increase* effect over time ("coining").

[2] All entries count as published that are published journal articles or book chapters.

2.6.2 Regression slopes

Unweighted regression slopes were calculated by using the primary data from the respective meta-analysis. Prior to calculating the regression slopes, the primary data from all studies which yielded a negative summary effect were multiplied by (-1), so that *negative* signs of regression coefficients show a *decline*, whereas *positive* regression coefficients show an *increase* effect over time ("coining").

2.7 Publication bias detection methods

As mentioned above, publication bias detection methods were applied to each eligible meta-analysis that provided primary data. The most common publication bias methods will be described in more detail below. Publication bias detection methods can be divided into several "publication bias groups".

2.7.1 1st Group: Failsafe *N*

The Failsafe *N* (Rosenthal, 1979) was the first publication bias detection method which was developed. The idea behind it is to calculate the average number of studies that would be needed that the meta-analytic summary effect would become non-significant. If the Failsafe *N* is high, this would be interpreted as a robustness to publication bias because it would be unrealistic such a number of studies with non-significant effects exist (Becker, 2005).

The Failsafe *N* has been heavily criticised for its lack of statistical conceptualisation for the calculating formula and no solid interpretation of high numbers apart from a general rule of thumb (Becker, 2005). Furthermore, the number is primarily a statement regarding the significance of the meta-analytic summary effect. Given the fact that meta-analytic summary effects are most likely significant due to their higher power, their potential non-significant counterparts are negligible.

2.7.2 2nd Group: Visual inspection

Funnel plot

The easiest way for a researcher to investigate publication bias is by visually inspecting a funnel plot (Light et al., 1984). A funnel plot is a scatterplot between the size of

an effect size and its precision (e.g., the inverse standard error; Egger et al., 1997). In an ideal study without any publication bias, the scatterplot resembles a funnel shape with studies with high precision roughly scattering closer around the average and studies with lower precision scattering further away from the average. The funnel plot has been further developed, for example the contour-enhanced funnel plot (Peters et al., 2008).

The issue with using a funnel plot is that the interpretation of the plot is purely subjective and therefore not a suitable tool for bias interpretation (Terrin et al., 2005; Sterne et al., 2005).

2.7.3 3rd Group: Small study effects

"Small study effects" is the assumption that studies with small sample sizes systematically differ from studies with larger sample sizes (Siegel et al., 2022). For the visualisation of small study effects, the funnel plot is usually used und several statistical tests have been developed to statistically test its asymmetry.

Trim & Fill

The "Trim & Fill" method (Duval & Tweedie, 2000a, 2000b) is based on funnel plot asymmetry. When publication bias is present in the data, unprecise studies usually scatter around the far side of the funnel plot and "Trim & Fill" estimates how many studies you would need to "trim" so that the funnel plot to become symmetrical, in an iterative procedure (Duval, 2005). A new effect size estimate is then calculated with the symmetrical funnel plot and in the last step the trimmed studies and their statistical missing counterparts from the other side of the plot are "filled" back in. The final effect size estimate and its variance are calculated after these alternations but note that this corrected effect should only be seen as a sensitivity analysis.

"Trim & Fill" is a very intuitive method of interpretation, although it suffers from several shortcomings: As all methods that rely on funnel plot asymmetry, it suffers from unprecise estimates when heterogeneity is present (Terrin et al., 2003) or overestimation of small study effects not caused by publication bias (van Aert et al., 2019).

Egger's regression test

Sterne and Egger's (2005) regression test is a weighted regression of the effect size on its precision (i.e., the standard error). A significant intercept (two-sided *p*-value of .10) is an indicator for publication bias.

Begg and Mazumdar's rank correlation test

Begg and Mazumdar's (1994) rank correlation test is a correlation between standardised effect sizes and their respective variances but based on an adjusted Kendall's rank correlation. A significant rank correlation (two-sided *p*-value of .10) is indicative for publication bias.

Both Egger's regression test and Begg and Mazumdar's rank correlation have been criticized for suffering from low statistical power (Borenstein et al., 2009) and increased false-positive rates with increasing heterogeneity (Ioannidis & Trikalinos, 2007).

PET-PEESE

PET-PEESE (Stanley & Doucouliagos, 2014) is a further development of the Egger's regression test. The idea is to get a meta-analytic summary effect which is adjusted for small study effects. The *precision estimate test* (PET) is again a weighted regression of the effect size on its precision (the standard error). A significant intercept ($p < .05$) indicates publication bias. Since PET suffers from bias as well, it is possible to calculate an estimate using the variance (instead of the standard error) as the predictor of the effect size (*precision-effect estimate with standard error*, PEESE). If PET is significant, PET is interpreted. If PET is non-significant, PEESE will be interpreted as the bias-corrected intercept. If PEESE is non-significant, no bias is present.

As PET-PEESE is very similar to the Egger's Regression test, it suffers from many of the same problems, i.e., low study power and unprecise estimates when heterogeneity is present (Alinaghi & Reed, 2018).

2.7.4 4th Group: Methods based on *p*-values

Methods that are based on *p*-values work with the assumption that under the null hypothesis the *p*-value of studies are uniformly distributed. If there is an effect present, the distribution will become skewed. These methods (except *p*-uniform*) only work with significant *p*-values under the assumption that all significant *p*-values should all have the same probability of being published. The idea is to investigate the "evidential value" of studies. That means the idea is to investigate whether a true effect exists in a given dataset which does not stem from selective reporting or misuse of statistical methods (e.g., *p*-hacking).

p-curve

P-curve assesses the distribution of *p*-values in a meta-analysis. In the case of a "true" effect the *p*-curve should be right-skewed and there should be more smaller ($p < .01$) than larger ($p < .05$) *p*-values, as we would expect the true effect to be revealed on a higher significance level that is closer to .01 than .05. If there is no effect, the *p*-curve should be flat with the *p*-values distributed uniformly. In the case of a non-true effect (which might stem, for example, from QRPs), the *p*-curve is left-skewed with more larger than smaller *p*-values (Simonsohn et al., 2014). When *p*-hacking is present, the *p*-values usually cumulate around the significance threshold of $p < .05$, reaching just about nominal significance.

The right skew of *p*-curve can be statistically tested, either by binomial test or by a continuous one. In the binomial test, the proportion of observed *p*-values below .025 is compared to the assumed proportions of *p*-values below .025 if the null hypothesis was valid. A significant result indicates a right skew (one-sided *p*-value).
The continuous test consists of two tests. First, it is possible to calculate the so-called *pp*-values (= the probability to get a *p*-value that is as extreme as the initial one). The *pp*-values are aggregated using Stouffer's method (Stouffer et al., 1949; Simonsohn et al., 2015) to calculate a so-called "full *p*-curve" (all *p*-values <.05) and a "half *p*-curve" (all *p*-values <.025). The idea is to show that larger biases can be observed not only in the full *p*-curve but also in the half *p*-curve with the smaller *p*-values. When the *p*-curves are right skewed, this is seen as "evidential value".
The second test can be used to assess whether the included studies suffer from low power. This is done by calculating a set of studies with an average power of 33% (Simonsohn et al., 2014, 2015) and then comparing the respective *p*-curves with the actual set. If the original *p*-curve is flatter than the one calculated with 33% power, this can be seen as an indicator that the given evidence is not robust enough that one could speak of a non-trivial effect in the sample. Furthermore, it is possible to calculate an adjusted effect size estimate with *p*-curve. In an iterative procedure the effect size is estimated that would render the *pp*-values into a uniform distribution.

p-uniform

P-uniform is similar to *p*-curve in terms of the statistical principles and hypothesis testing and it is also possible to test for inflated effects by comparing the distribution of the *p*-values. Again, a uniform distribution of significant *p*-values is assumed but in contrast to *p*-curve, which tests for skewness, *p*-uniform tests for the uniformness of the distribution. The

idea is to check for a significant deviation between the transformed *p*-values and the uniform distribution. The significance test is based on a Fishers test which checks for the uniform distribution with $p < .10$ as level of significance. If there is a significant deviation, this can be seen as an indication for bias. Additionally, it is possible to calculate an adjusted effect size estimate, again with an iterative approach that would render the *p*-values into a uniform distribution.

Both *p*-curve and *p*-uniform have been criticised for their use of only significant effect sizes and not using the full scale of effect sizes available (van Aert & van Assen, 2018). Moreover, in cases of heterogeneity, *p*-curve and *p*-uniform tend to overestimate effect sizes (McShane et al., 2016; van Aert et al., 2016) or to underestimate in cases where *p*-hacking is present (van Aert et al., 2016).

p-uniform*

P-uniform* was developed to overcome the criticisms mentioned above. Firstly, *p*-uniform* includes both significant and non-significant effect sizes. In the case of between-studies heterogeneity the estimation for the adjusted effect size was improved and therefore allows to investigate the extent of unobserved heterogeneity (van Aert & van Assen, 2018). In the present analysis, for *p*-uniform and *p*-uniform* only primary effect sizes with signs that match the summary effect sign will be included (i.e., Proteus effects were excluded).

Test of excess significance

The test of excess significance (Ioannidis & Trikalinos, 2007) is used to check whether too many significant effects appear in a dataset. The number of observed significant effects is compared to the number of significant effects that would be expected given the average power of the studies in the dataset. The expected effects are calculated a-priori with a post-hoc power analysis of the respective studies. The comparison is done with a χ^2 test with $p < .10$ as level of significance. If the χ^2 test is significant, and there are more observed than expected significant effects, this can be seen as an indicator for publication bias.

The test of excess of significance has been criticized for poor performance when heterogeneity is present (Ioannidis & Trikalinos, 2007; Renkewitz & Keiner, 2019). It also struggles with assessing the magnitude of bias present in a meta-analytic dataset.

2.7.6 5th Group: Selection Model approaches

Vevea and Woods' Selection Model

Selection models were developed to adjust for publication bias by combining two models for a bias-adjusted meta-analytic estimate. The "selection model" investigates the likelihood of an effect size being published (mostly through measures like their significance value, i.e., *p*-value) and assigns certain weights to this likelihood. Effect sizes that are unlikely to be published get assigned a larger weight (van Aert & van Assen, 2018). The "effect size model" estimates the effect size distribution under the assumption that publication bias is not present.

Vevea and Woods (2005) designed a selection model that uses pre-defined weights. The meta-analysts can either compute their own weight functions or choose between four pre-existing ones ("moderate one-tailed selection", "severe one-tailed selection", "moderate two-tailed selection", and "severe two-tailed selection") – according to certain assumptions about their data (e.g., the amount of bias). If no publication bias is present, the adjusted estimate should not differ to a large extent from the meta-analytic estimate from the data (Vevea et al., 2019).

Selection models are not often applied, as they are not intuitive to use. However, the implementation of such models has been recommended because they outperform most other publication bias detection methods (Carter et al., 2019; McShane et al., 2016).

2.8. Triangulation

Since none of the publication bias detection methods is without flaws, meta-analysts have recommended the use of *triangulation* (Cooper et al., 2019; Kepes et al., 2022, Field et al., 2020). This means one should test for publication bias not with a single method but with numerous ones – in the best case with different tests from different "publication bias groups" (e.g., one test from the group of small study effects, one that relies on *p*-values and a selection model). This way the publication bias methods can compensate for each other's weaknesses and researchers can make a more differentiated statement whether publication bias is present or not.

For this study, we used nine meta-analytic bias detection methods, deriving from three different "groups" (see 2.10.1 "The Meta-level").

2.9. Inference criteria

Table 1 shows the defined thresholds that indicate publication bias for the respective publication bias detection methods. For more detailed information regarding the thresholds, see Siegel et al. (2022).

2.10 Statistical analyses

The analysis takes place on two levels: The *Meta-level* and the *Meta-Meta-level*.

Table 1. *Thresholds for bias indication for each publication bias detection method.*

Method	Threshold
Trim & Fill	Trim & Fill indicates that studies are missing
Egger's Regression	$p < .10$
Rank Correlation	$p < .10$
PET-PEESE	Difference between the summary effect and the adjusted effect exceeds 20%
p-curve	For half *p*-curves: $p < .05$, for half and full *p*-curves: $p < .10$
p-uniform	$p < .10$ (for one-sided test)
p-uniform* [a]	$p < .10$ (for one-sided test)
Excess of significance	$p < .10$ & number of observed ES > number of expected ES
Selection models	Difference between the summary effect and the adjusted effect exceeds 20%

Note. ES = effect size.

[a] Thresholds in accordance with Siegel et al. (2022), except for *p*-uniform*, as the planned publication bias test was not possible.

2.10.1 The Meta-level

The focus of the Meta-level lies on the primary data of the included meta-analyses. Calculations are carried out for each meta-analysis separately.

The calculations here include:

- Calculating the indices, i.e., the crude differences and the regression slopes.
- Recalculating the summary effect sizes of the meta-analyses using REML (restricted maximum likelihood).

The following publication bias detection methods are used to investigate and detect possible publication bias:

- Trim and Fill (Duval & Tweedie, 2000a, 2000b)
- Rank Correlation Test (Begg & Mazumdar, 1994)

- Egger's Regression Test (Sterne & Egger, 2005)
- PET-PEESE (Stanley & Doucouliagos, 2014)
- Selection model (Vevea & Woods, 2005)
- Test of Excess Significance (Ioannidis & Trikalinos, 2007)
- *p*-curve (Simonsohn et al., 2014a, 2014b)
- *p*-uniform (van Aert et al., 2016)
- *p*-uniform* (van Aert & van Assen, 2018)

2.10.2 The Meta-Meta-level

The focus of the Meta-Meta level lies on the combination of the indices (crude differences and regression slopes) from the Meta-level, which serve as the primary data here. The goal is to investigate the effect changes over time and to see if evidence of effect misestimation can be found. Calculations were carried out for all included meta-analyses and their respective indices.

The calculations here include:

- Average effect sizes for the crude differences (CD) and the regression slopes (RS) were calculated using REMLs (restricted maximum likelihood).
- Subgroup analyses were calculated using mixed-effect models to investigate the occurrence and strength of decline and increase effects and to check for significant differences between declines and increases. This was done for the crude differences and for the regression slopes separately. The subgroup analysis was done first with the complete datasets (published and grey literature) and then with results from published studies within the meta-analysis only.
- A second subgroup analysis was calculated to investigate whether a difference in effect size metric influences crude differences. Studies based on mean comparisons (Cohens *d*, Hedges *g*) were compared to correlational studies (Pearson *r*, Fishers *z*).
- Influences on heterogeneity were calculated using single linear meta-regressions, thereby investigating the influence of certain study characteristics (sample numbers (*k*), participant numbers (total *N*) and average participant numbers (average *N*)) on the heterogeneity estimates (I^2 and tau^2) within a sample.
- To examine the association of certain study characteristics (regarding the initial study and the meta-analysis) and measures of effect changes over time (CD and RS), single and theory-guided hierarchical multiple regressions were calculated.

Multicollinearity will be assessed using the *Variance Inflation Factor* (VIF). If the VIFs reach a value higher than four during the regression, the predictor with the highest value will be dropped out of the analysis. This is repeated until all remaining VIFs are below four.

The meta-regression is done first with the complete datasets (published and grey literature) and then with results from published studies within the meta-analysis only.

The predictors for the hierarchical multiple regressions are:

Initial study characteristics:

Model 1 (dependent variable = CDs) & *Model 3* (dependent variable = RS)

- i.) Initial study absolute effect size
- ii.) Initial study sample size
- iii.) Initial study publication year
- iv.) Initial study annual citation numbers in the Web of Science
- v.) Initial study journal impact factor

Meta-analytic characteristics:

Model 2 (dependent variable = CDs) & *Model 4* (dependent variable = RS)

- i.) Absolute summary effect size
- ii.) Number of included samples (*k*)
- iii.) Covered time span of the meta-analysis
- iv.) Type of included primary studies in a meta-analysis
  (not included in the second calculation (= only published))
- v.) Publication year of the meta-analysis

- To investigate the assumption that publication bias detection methods will detect more bias if only published studies are included (rather than including grey literature as well), χ^2 tests were calculated. The crosstab includes whether a study contains all studies ("published and grey literature") or only published studies ("pub only") and whether the bias detection method indicated the presence of publication bias ("bias") or not ("no bias"). This was done for each publication bias detection method

individually. If the requirement of $n < 5$ per cell is not given, Fishers Exact Test was calculated.

2.11 Awareness of temporal effect changes in the relevant scientific community

In an exploratory study we sought to investigate whether the scientific community of clinical child psychology is aware of the mechanisms of temporal effect changes, the possible reasons, and outcomes.

To this end, an online survey was sent out to all researchers who have published in the *JCPP* via email. Based on their successful publication in the journal, we considered them experts in the field and therefore our target population. In the survey, which was conducted with Qualtrics, researchers were asked four short questions and afterwards certain demographics were assessed.

The four questions were:

1.) Prediction of direction: Estimate whether the overall absolute effect sizes (regardless of the research question) in the JCPP have been getting larger, smaller or have stayed the same magnitude over time (from the 1960ies to today)? [3]

2.) Open Question: Please explain in a sentence or two, why you think this effect change might be happening.

3.) Magnitude ES (initial studies): Estimate the average overall effect size from the studies *from the 1960ies* using Pearson *r*.

4.) Magnitude ES (now): Estimate the average overall effect size from the *studies from the last two* years using Pearson *r*.

The demographics assessed were:

- Occupation (graduate student, post-doc or non-tenured professor, professor – tenure track, professor – tenured, other)
- Age in ranges (18-25, 26-35, 36-45, 46-55, 56-65, 66+)

[3] The 1960s were named as the starting point for the temporal effect changes here, since the oldest initial study in our sample was published in 1963.

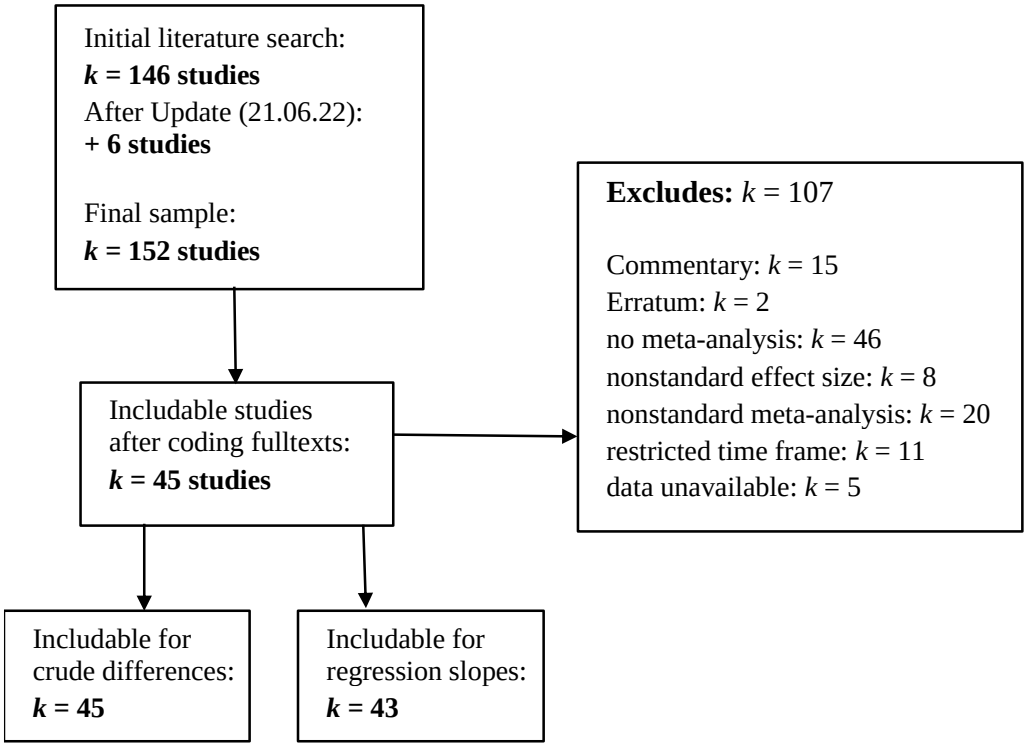
- Consideration of oneself as an expert for clinical child psychology and psychiatry (Yes, No, A bit)
- PhD or MD (PhD, MD, none)

3. Results

3.1 Literature Search

The initial literature search yielded 146 results plus additional six results after the update. Eight studies required further information from the authors and in summary, 45 studies were included. Two authors provided data, whereas 107 studies were excluded. Figure 1 shows the flowchart for the inclusion and exclusion of studies.

Figure 1. *Flowchart depicting the selection process for the meta-analyses.*



3.2 Coding

Cohen's kappa for agreement between the two coders was high with $\kappa = .85$ for the inclusion and exclusion of studies. Discrepancies were discussed and resolved. In case of no agreement, a third researcher was consulted, and findings were discussed until consensus was reached. Forty-five studies allowed the calculation of crude differences, and the calculation of regression slopes was possible in 43 studies.

Note that Proteus effects were excluded in further analyses, as there is no general consensus about how to handle Proteus effects. They can either be seen as an extreme case of decline effect (and should therefore consequently be added to the declines) or there might be some independent yet unknown mechanism, which causes the effect misestimation (Pfeifer et al., 2011). With the exclusion of Proteus effects, 40 studies for the calculation of crude differences and 38 studies for the regression slopes remained.

3.3 Summary statistics

The final dataset with the data that were reported in all the 45 the meta-analyses included n = 1360 effect sizes (range: 3 – 239) and a total number of participants of N = 1.865.461. The global timespan covered 57 years, with the earliest study being published in 1963. When reanalysing the data provided from the meta-analyses, either by table, supplement, or author contact, we found n = 1255 included effect sizes (range: 3 – 237) and a total participant number of N = 590.991. The timespan and the earliest study were the same. The main summary effect over all meta-analytic summary effects was r = .206, p < .001, according to a random effects model.

All meta-analyses were calculated according to the Hedges and Olkin approach, no Hunter and Schmidt meta-analysis was found in the sample. Forty-one (91%) studies used a random effects model, and four (9%) studies used a fixed effect model.
Of all the included meta-analyses (k = 45), about 37% (k = 17) included grey literature and 73% (k = 33) included at least one publication bias detection method. Seventy percent (k = 23) included more than one publication bias detection method, on average 2.4 publication bias detection methods were used. From the studies which include at least one bias calculation method, 24% (k = 8) stated that they did find evidence for publication bias in their meta-analyses and 76% (k = 25) stated that they did not.

The most frequent method used to assess publication bias was the funnel plot (78%, k = 26), with the vast majority of studies using it. The next most frequent method was Egger's regression (54%, k = 18), followed by "Trim & Fill" (45%, k = 15) and Failsafe N (30%, k = 10). Four studies used the rank correlation test (12%), two studies (6%) p-curve, two (6%) used other publication bias detection methods (once Orwin's *FSN* and once a correlation between sample size and effect size), and one study (3%) used the direct comparison to assess publication bias.

The effect sizes were divided as follows: The vast majority of studies relied on Cohens *d* (44%, *n* = 20) and Hedges *g* (29%, *n* = 13). Both effect sizes compare means and can be mainly found in experimental studies. Effect sizes from correlational studies were lower in number, Pearson *r* is used by five (11%) studies and Fishers *z* by only two (5%) studies. Odds Ratios were used by five (11%) studies.

3.4 Average power

Since low study power was named as one of the reasons for publication bias and decline effect, the average, minimum and maximum power was calculated. Average power for all studies was *M* = 50.48%, Median power was *Md* = 47.14%, with a minimum power of *Min* = 6.30% and a maximum power of *Max* = 100%. This indicated that the average power is just about as high as random chance. The wide range of power was also reflected in the sample size span (with *k* = number of included studies per meta-analysis), which ranges from *k* = 3 to *k* = 237 (when the included meta-analyses were recalculated by the author, the reported sample size span was almost identical with *Min* = 3 and *Max* = 239).

3.5 Decline and increase effect

Table 2 shows the average decline and increase effects for crude differences and regression slopes.

Table 2. *Average declines, increases and Proteus effects for crude differences and regression slopes (complete dataset).*

Effect change	*k*	*r*	*p*
	crude differences		
Average overall	36	.0440	.2369
Decline effect	16	.2207	.0005***
Increase effect	19	.1140	.0112*
Proteus effect	5	.1868	.1039
	regression slopes		
Average overall	43	.0035	.0108*
Decline effect	27	.0078	<.0001***
Increase effect	16	.0038	.0030**

Note. *k* = number of effects, *r* = Pearson *r*. Absolute values are reported.
One case was omitted in the crude differences because the result of the subtraction of summary effect minus initial effect resulted in zero.
* *p* <.05, ** *p* < .01, *** *p* < .001

For crude differences, the subgroup analysis showed a statistically significant difference for declines vs increases, $Q(1) = 3.91$, $p = .048$. Results became non-significant when only published results were investigated (pub only: $k = 20$, decline pub: $k = 5$, increase pub: $k = 15$), $Q(1) = 0.35$, $p = .557$.
For regression slopes, subgroup analysis revealed a non-significant difference for declines vs increases, $Q(1) = 3.77$, $p = .052$. Results remained non-significant when only published results were investigated (pub only: $k = 17$, decline pub: $k = 17$, increase pub: $k = 10$), $Q(1) = 2.44$, $p = .119$.

The temporal effect change for every meta-analysis is illustrated in the lighting plot in Figure 2. On the X axis is the relative timespan (in years) plotted against the relative annual change in Fishers *z* on the Y axis. Each line represents one meta-analysis, each bend is a datapoint. Blue lines represent declines, whereas pink lines represent increases. The ratio of declines to increases is 1.33:1.

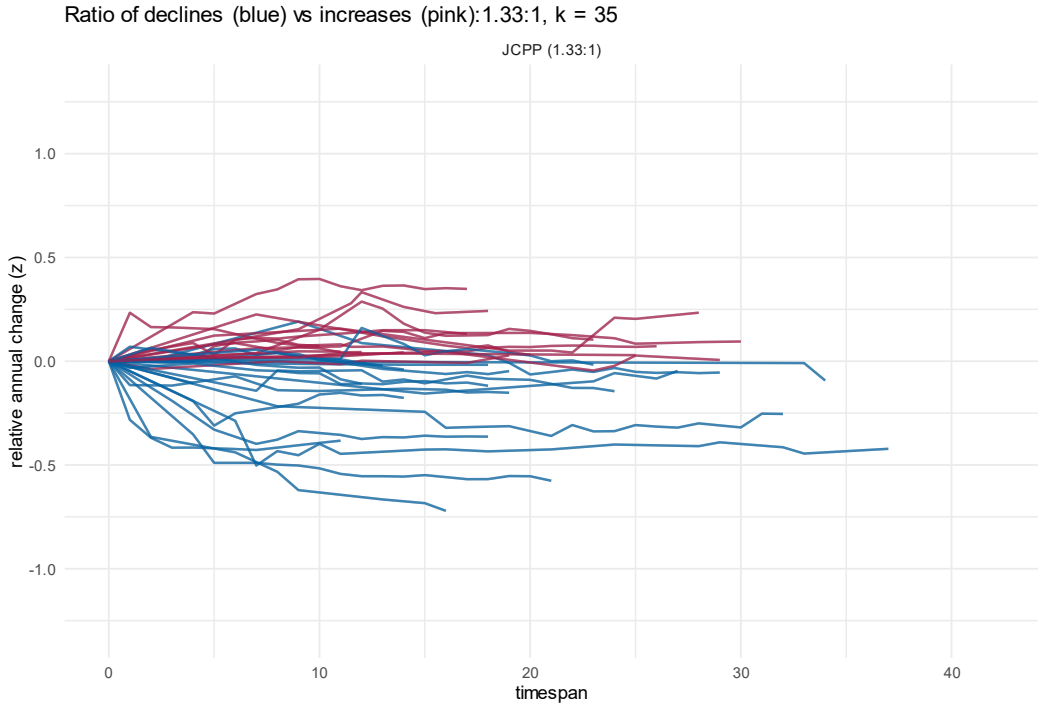


Figure 2. *Lightning plot illustrating declining (blue) and increasing (pink) effects over the relative timespan.*

3.6 Differences in effect size metric

Subgroup analyses revealed that meta-analyses that were based on correlational data differed statistically significantly from meta-analyses based on means, $Q(1) = 4.73$, $p = .030$. The average summary effect for means was declining (average $r = .067$, $p = .143$) whereas the average summary effect for correlational data was increasing (average $r = .121$, $p = .099$).

Effect sizes were small and non-significant. Note that the majority of studies investigated measures based on means (k = 27) and only few investigated correlations (k = 5).

3.7. Meta-regression I^2 and tau^2

Table 3 shows the characteristic values for the meta-regressions on I^2 and tau^2. Total N was significant in both cases and the b's show a positive association. Average N showed a significant influence on I^2 with a positive association. The sample number was non-significant in both cases.

Table 3. *Results from the meta-regression for I^2 and tau^2.*

	I^2			tau^2		
Predictor	*b*	*p*	R^2	*b*	*p*	R^2
Sample number (*k*)	< -.0001	.0983	0.00%	.0001	.0906	11.63%
Total *N*	< .0001	.0001***	50.29%	< .0001	.0078**	12.71%
Average *N*	.0002	.0007***	42.77%	< -.0001	.6187	0.00%

Note. ** p < .01, *** *p* < .001

3.8 Multiple hierarchical regressions

Multiple hierarchical regressions were calculated for various predictors regarding initial study characteristics and meta-analytic characteristics. The hierarchical regressions were calculated for crude differences and regression slopes separately. First, the analyses were carried out with the complete dataset (published and unpublished literature) and then for published literature only.

Variance Inflation Factors (VIF's) were calculated to check for multicollinearity of predictors. If multicollinearity is high, type II error is inflated, making the results of the regression unreliable. All VIF's were constantly below four, meaning there was very little evidence for multicollinearity of predictors.

As some predictors had missing values, first a full model and then a reduced model was calculated for each Model and Step to ensure that the reduction to a dataset without missing values did not alter results. Only minor differences between the results of the full model and the reduced model emerged.

For all analyses regarding the crude difference, the values that were reported in the meta-analyses were used. For all analyses regarding the regression slopes, the reanalysed values from the recalculation of the meta-analyses were used.

3.8.1 Complete Dataset

Table 4 shows the analysis using the complete dataset for crude differences. They provided the following model fits: For Model 1, Step 1 with the intercept and the absolute initial effect size provided the best model fit with $R^2 = 97.33\%$, $\chi^2(1) = 27.02$, $p < .001$.

Table 4. *Results for the best model fit for the multiple hierarchical regression of crude differences with the complete dataset.*

Model 1 (complete dataset): crude differences & initial study characteristics		
Predictor	*b*	*p*
Absolute initial ES	0.7345	< .0001***

Model 2 (complete dataset): crude differences & meta-analytic characteristics		
Predictor	*b*	*p*
Absolute summary ES	-0.4494	.1320
k of ES	-0.0013	.0900
Covered time span	-0.0010	.8452
Unpublished studies	0.2809	.0029*

Note. ES = effect size.
* p < .05, *** p < .001

Table 5. *Results for the best model fit for the multiple hierarchical regression of regression slopes with the complete dataset.*

Model 3 (complete dataset): regression slopes & initial study characteristics		
Predictors	*b*	*p*
Absolute ES	-0.0042	.5409
n initial study	< 0.0001	.0108*

Model 4 (complete dataset): regression slopes & meta-analytic characteristics		
Predictors	*b*	*p*
Absolute summary ES	-0.0016	.8500
k of ES	-0.0001	< .0001 ***
Covered time span	< 0.0001	.5838
Unpublished studies	0.0006	.7603
Publication year	0.0007	< .0001***

Note. ES = effect size.
* *p* < .05, *** *p* < .001

For Model 2, Step 4 with the intercept and the absolute meta-analytic summary effect size and the number of effect sizes in a meta-analytic study (*k*) and the covered time span and the occurrence of unpublished studies provided the best model fit with R^2 = 48.68%, $\chi^2(4)$ = 09.84, p = .043.

Table 5 shows the analysis using the complete dataset for regression slopes. They provided the following model fits: For Model 3, Step 2 with the intercept and the absolute initial effect size and the *n* of the initial study provided the best model fit with R^2 = 38.13%, $\chi^2(2)$ = 08.26, p = .016. For Model 4, Step 5 with intercept and the absolute meta-analytic summary effect size and the number of effect sizes in a meta-analytic study (*k*) and the covered time span and the occurrence of unpublished studies and the publication year provided the best model fit with R^2 = 95.19%, $\chi^2(5)$ = 39.40, p <.001.

3.8.2 Published data only

Table 6. *Results for the best model fit for the multiple hierarchical regression of crude differences with published data only.*

Model 1 (published only): crude differences & initial study characteristics		
Predictor	*b*	*p*
No significant predictor	-	-

Model 2 (published only): crude differences & meta-analytic characteristics		
Predictor	*b*	*p*
Absolute summary ES	-0.7649	.0167*

Note. ES = effect size.

No Step 4 (unpublished studies) in Model 2, as only published studies are investigated.

* p < .05

When re-running the analyses with published data only for crude differences some changes emerged, as can be seen in Table 6. For Model 1, the absolute initial effect size had no influence anymore when only published data was investigated. No predictor showed any influence, R^2 = 80.20%, $\chi^2(1)$ = 02.90, p = .089.

For Model 2, Step 1 with only the intercept and the absolute meta-analytic summary effect provided the best model fit now R^2 = 100%, $\chi^2(2)$ = 06.41, p = .041. The other predictors (number of effect sizes in a meta-analytic study (*k*), covered time span) had no influence.

When re-running the analyses with published data only for regression slopes, the model fits did not change. For Model 3, Step 2 with the intercept and the absolute initial effect size and the *n* of the initial study still provided the best model fit with R^2 = 32.96%, $\chi^2(2)$ = 07.09, p = .030. For Model 4, Step 4 with the intercept and the absolute meta-analytic summary effect size and the number of effect sizes in a meta-analytic study (*k*) and the covered time span and the publication year also still provided the best model fit with R^2 = 98.08%, $\chi^2(4)$ = 46.39, $p < .001$ (note that Step 4 is the equivalent of Step 5 with the complete dataset, as only published studies are investigated now and the fourth predictor is "unpublished studies"). See Table 7 for *b*'s and *p*-values.

Table 7. *Results for the best model fit for the multiple hierarchical regression of regression slopes with published data only.*

Model 3 (published only): regression slopes & initial study characteristics		
Predictors	*b*	*p*
Absolute ES	-0.0046	.5189
n initial study	-0.0005	.0205*
Model 4 (published only): regression slopes & meta-analytic characteristics		
Predictors	*b*	*p*
Absolute summary ES	-0.0027	.7434
k of ES	-0.0001	< .0001 ***
Covered time span	< 0.0001	.7884
Publication year	0.0008	< .0001 ***

Note. ES = effect size.

No Step 4 (unpublished studies) in Model 4, as only published studies are investigated.

* $p < .05$, *** $p < .001$

The tables with the complete multiple hierarchical regression can be found in the supplemental material S2.

3.9 Single Regressions

Single regressions for each predictor from above were calculated again for crude differences and regression slopes. Calculations were once done with the complete dataset and once with published studies only.

3.9.1 Complete Dataset

Table 8 shows the results for the crude differences with the complete dataset. In single regressions, the *initial effect size* and the *meta-analytic publication status* were the only significant predictors.

Table 8. *Single regressions over crude differences with the complete dataset.*

Crude differences (complete dataset)			
Predictor	*b*	*p*	R^2
Initial absolute ES	0.7318	< .0001***	100%
Initial study *N*	-0.0002	.1471	16.44%
Initial publication year	-0.0024	.6010	0.00%
Initial number of citations	0.0021	.7086	0.00%
Initial Impact Factor	-0.0191	.1355	13.78%
MA absolute summary ES	-0.1908	.5513	0.00%
MA *k* of ES	< -0.0001	.9544	0.00%
MA publication year	-0.0028	.7412	0.00%
MA covered timespan	0.0034	.5263	0.00%
MA publication status	0.1933	.0250*	19.31%

Note. ES = effect size, MA = meta-analytic

* $p < .05$, *** $p < .001$

This matches the results from Model 1, where also the initial effect size was the significant predictor apart from the intercept. The *b*'s are almost identical as were the significance levels ($b_{\text{multi reg}} = 0.735$, $p < .001$ vs. $b_{\text{single reg}} = 0.732$, $p < .001$). Moreover, the meta-analytic publication status aligns with Model 2 from the multiple hierarchical regression, as Step 4 indicated the publication status. However, in contrast to the initial effect size, *b*'s and significance levels differed here ($b_{\text{multi reg}} = 0.281$, $p = .003$ vs. $b_{\text{single reg}} = 0.193$, $p = .025$).

When looking at the regression slopes in Table 9, *the initial n, the meta-analytic sample size,* and the *publication year* were significant predictors. This only partially aligns with the multiple hierarchical regression, as Step 2 in Model 3 also included the initial *n* ($b_{\text{multi reg}} = <0.001$, $p = 011$), but apart from that, no predictors overlap here.

3.9.2 Published data only

When looking at the analyses with only published data for the crude differences in Table 10, the *meta-analytic summary effect* was the only significant predictor.

Table 9. *Single regressions over regression slopes with the complete dataset.*

Regression slopes (complete dataset)			
Predictor	*b*	*p*	R^2
Initial absolute ES	-0.0060	.3124	01.83%
Initial study *N*	< 0.0001	.0058**	31.13%
Initial publication year	0.0001	.3699	2.26%
Initial number of citations	< 0.0001	.7189	0.00%
Initial Impact Factor	< 0.0001	.9851	0.00%
MA absolute summary ES	-0.0076	.4977	0.00%
MA *k* of ES	< -0.0000	.0295*	20.81%
MA publication year	0.0006	.0045**	37.02%
MA covered timespan	< 0.0001	.6601	0.00%
MA publication status	< 0.0001	.9763	0.00%

Note. ES = effect size, MA = meta-analytic

* $p < .05$, ** $p < .01$

Table 10. *Single regressions over crude differences with published data only.*

Crude differences (published only)			
Predictor	*b*	*p*	R^2
Initial absolute ES	0.6433	.1022	0.00%
Initial study *N*	-0.0002	.2195	0.00%
Initial publication year	-0.0017	.7208	0.00%
Initial number of citations	-0.0003	.9575	0.00%
Initial Impact Factor	-0.0197	.1435	0.00%
MA absolute summary ES	-0.7649	.0167*	100%
MA *k* of ES	-0.0036	.2970	48.14%
MA publication year	-0.0032	.7691	0.00%
MA covered timespan	0.0016	.7679	0.00%

Note. No publication status as predictor as only published studies are investigated.

ES = effect size, MA = meta-analytic

* $p < .05$

Compared to the predictors from the complete dataset from the single regressions, the meta-analytic summary effect did not play a role there. However, it aligns again with Model 2 from the multiple hierarchical regression with published data, where Step 1 with the intercept and the meta-analytic summary effect also provided the best model fit with almost identical *b*'s ($b_{\text{multi reg}}$ = -0.765, p = .017 vs. $b_{\text{single reg}}$ = -0.765, p = .017).

Table 11. *Single regressions over regression slopes with published data only.*

Regression slopes (published only)			
Predictor	*b*	*p*	R^2
Initial absolute ES	-0.0063	0.3139	01.49%
Initial study *N*	< 0.0001	0.0157*	19.23%
Initial publication year	0.0001	0.3692	03.05%
Initial number of citations	< 0.0001	0.7399	0.00%
Initial Impact Factor	< 0.0001	0.9979	0.00%
MA absolute summary ES	-0.0069	0.5555	0.00%
MA *k* of ES	< -0.0001	0.0321*	19.16%
MA publication year	0.0062	0.0040**	40.46%
MA covered timespan	< 0.0001	0.6819	0.00%

Note. No publication status as predictor as only published studies are investigated.
ES = effect size, MA = meta-analytic
* p < .05, ** p < .01

For the regression slopes in the published dataset, the significant predictors were *the initial study n, the meta-analytic sample size* and *the publication year* as shown in Table 11. These predictors were identical to the predictors when looking at the complete dataset with almost identical *b*'s. Compared to the multiple hierarchical regression where Model 3, Step 2 with the intercept and the initial study *n* provided the best model fit, *the initial study n* was a significant predictor here as well. The *meta-analytic sample size and publication year* also were significant predictors in Model 4, where Step 4 provided the best model fit and that included both meta-*analytic sample size and publication year* as significant predictors.

3.10. Publication bias detection methods

Nine different publication bias detection methods were calculated, and the indication of bias was interpreted according to the thresholds from Siegel et al. (2022, see Table 1 above). Table 12 shows the frequency of bias occurrence for each method.

Considering the triangulation, the small-study-effects-methods indicated bias the most often, followed by the selection models. Methods based on *p*-values all showed very similar occurrences of bias. When investigating only published data (values in parentheses), the bias indications were around two times as high for small-study-effects-methods and the selection models. For methods based on *p*-values, *p*-uniform* showed a bias indication even three times higher for published data. Excess of significance and *p*-uniform had similar bias indications, about two times higher for published data. Only *p*-curve showed less bias indication for published data (one instead of two). Bias was indicated in all different "publication bias groups", although the range was quite wide (1 – 27).

Table 12. *Frequency of bias indication per publication bias detection method.*

Publication bias method	Frequency of bias
PET-PEESE	11 (27)
Egger's Regression	9 (17)
Rank Correlation	9 (16)
Trim & Fill	8 (19)
Selection models	9 (21)
p-uniform*	5 (18)
Excess of Significance	3 (5)
p-curve	2 (1)
p-uniform	1 (4)

Note. Values in parentheses represent the frequency of bias indication when only published data was investigated.

3.11 Influence of grey literature

To investigate whether grey literature influences the occurrence of publication bias, χ^2 tests or Fishers Exact tests were calculated. Table 13 shows the results.

None of the tests used displayed significance. For the detection of publication bias, the inclusion of grey literature did not seem to make a difference.

3.12 Survey „Awareness of decline effect in clinical child research"

The survey was sent out on 15.11.22 and was closed by 31.12.22. In total, 2181 researchers were contacted. 136 (6%) responses were reported and 115 (5%) remained after removing all entries that were left blank.

Table 13. χ^2 *test and Fishers exact test for all publication bias detection methods*

Publication bias method	χ^2(df)	*p*
	χ^2 *test*	
Trim & Fill	0.1588(1)	.6903
Egger's Regression	1.3218(1)	.2503
Rank Correlation	1.7314(1)	.1882
PET-PEESE	0.1022(1)	.7492
p-uniform*	1.0358(1)	.3088
	Fishers exact test	
Excess of significance	-	.8390
p-curve	-	.9761
p-uniform	-	.5438

Note. For Fishers exact test, only the *p*-value is interpreted.

Question 1: Prediction of effect change direction (*n* = 115)

Q1: "Estimate whether the overall absolute effect sizes (regardless of the research question) in the JCPP have been getting larger, smaller or have stayed the same magnitude over time (from the 1960ies to today)?"

Figure 3 shows the distribution of the different predictions of effect change over time.

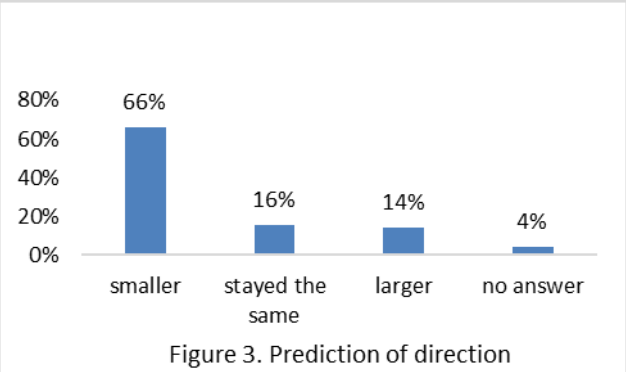


Figure 3. Prediction of direction

Seventy-six (66%) researchers predicted smaller effect sizes over time, while 18 (16%) researchers estimated that effect sizes stay the same over time. Larger effect sizes over time were predicted by 16 (14%) researchers. Five (4%) researchers provided no answer for this question.

Question 2: Open Question (n = 105)

Q2: “Please explain in a sentence or two, why you think this effect change might be happening.”

The responses to this question were recorded and clustered into categories. Figure 4 shows their distribution. The majority of researchers attributed the changing of effects to the fact that study designs and science itself have improved over time. As reasoning for this assumption, the researchers mentioned “more precise measurements”, “more scientific rigor” and “better diagnostic instruments”. The next most frequent the explanation was the publication bias itself, where researchers attributed changing effect sizes to publication bias related mechanisms. The argument “samples are larger” was assigned its own category. Note that, larger samples have been named as a possible reason for declining effects only. “Other” explanations included reasons such as “secular trends”, “metrics” or “no major breakthroughs”. Clinical explanations included “changes in diagnostics”, “smaller samples with more impaired patients” or “more active comparisons”. Responses were coded as invalid when they provided statements such as “Not sure I understand the questions […]” or “I’m honestly just guessing”.

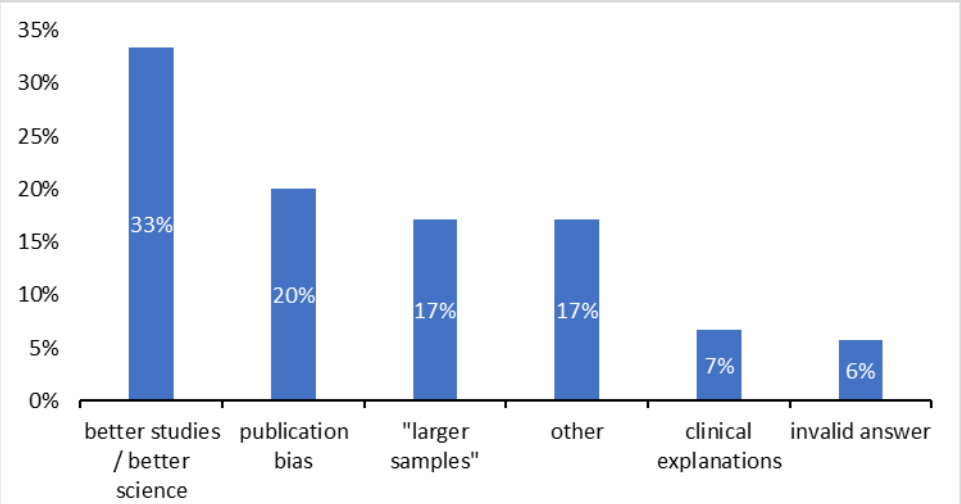


Figure 4. Answers to the open question divided into categories.

Question 3: Estimation of initial effect size (n = 82)

Q3: “Estimate the average overall effect size from the studies from the 1960ies using Pearson r.”

After removing impossible values (i.e., values greater than 1), the average effect size for the overall initial effect size was r = .436.

Question 4: Estimation of current effect size (n = 88)

Q4: “Estimate the average overall effect size from the studies from the last two years using Pearson *r*.”

After removing impossible values (i.e., values greater than 1), the average effect size for the overall current effect size was r = .321.

Demographics:

Occupation (n = 108):

The majority of responses came from tenured professors (n = 58, 54%) and post-docs / non-tenured professors (n = 21, 19%). Tenure track professors and “other” occupations (the most frequent response given here was “emeritus”) made up 19% (n = 13) of responses respectively. Two percent were grad students (n = 2) and 1% (n = 1) did not provide a response.

Age (n = 108):

Age could be indicated in different ranges. Most researchers were aged between 36-45 years (n = 37, 34%) at the time of the survey. The next most frequent age group was 66+ years with 23% (n = 25) and 46-55 years with 19% (n = 21). Eleven percent (n = 12) of the researchers were aged between 26-35 years and 55-56 years respectively. Only one participant (1%) was aged between 18-25 years.

PhD / MD (n = 108):

Eighty-three percent (n = 90) of researchers held a PhD degree, while 12% (n = 13) of researchers held an MD degree. Five persons (5%) held neither a PhD nor an MD degree.

Consideration of expertise (n = 108):

More than half (n = 72, 67%) of the researchers who completed the survey considered themselves experts in clinical child psychology. Thirty percent (n = 32) considered themselves “a bit” of an expert. Four participants (4%) did not consider themselves experts in clinical child psychology.

3.13 Exploratory analyses

To answer some of the assumptions that were raised through the survey, two additional exploratory analyses were conducted: First, as many researchers reported “larger

samples" as the main reason for effect change, we sought to investigate whether sample sizes really did increase over time. To this end, a spearman rank correlation was conducted for year and sample size. The result was a significant, positive correlation of *rho* = .211, $p < .001$. Second, we asked researchers to predict the current average overall effect size for the 1960s and now. The current average overall effect was also something that we can directly compare with our empirical data. While the average estimation the researchers gave was $r = .321$, our overall main effect from all the included meta-analyses was $r = .206$, $p < .001$.

4. Discussion

The aim of this study was to investigate whether temporal effect changes such as decline effect and increase effect can be found in the scientific literature of clinical child psychology and if publication bias can be identified as one of the drivers for it.

Decline effects and increase effects were indeed found (see Table 2). Contrary to previous findings (Pietschnig et al., 2019; Siegel et al., 2022), we found more increasing ($k = 19$) than declining ($k = 16$) effects for the crude differences in the complete dataset, but the effect misestimation was still substantially larger in the decline effects with almost two times the size of the increasing effects. When investigating decline and increase effects in the dataset with only published data from the meta-analyses, we found more than twice as many increases ($k = 14$) than declines ($k = 5$). It seems that the increase effects mainly stem from the published effect sizes, as they seem to cluster in the dataset with the published data only compared to the complete dataset with published and unpublished studies. In contrast to previous literature (Pietschnig et al., 2019, Siegel et al., 2022), these effects are increasing instead of decreasing. This is a new finding, and although no conceptual reason can be provided here, it could be seen as another indicator of effect size misestimation in published literature.

Regression slopes were calculated from the primary data and effect misestimation was investigated. For regression slopes we found substantially more declines ($k = 27$) than increases ($k = 16$) of similar sizes in the complete dataset. When investigating the dataset with published data only, effect misestimations revealed a consistent picture, with again more declining ($k = 17$) than increasing effects ($k = 10$). This aligns with the previous literature (Pietschnig et al., 2019, Siegel et al., 2022), which also showed an occurrence of more declining than increasing effects.

The effect misestimation found here can be interpreted as a genuine decline effect, which can be found in previous literature as well (Pietschnig et al., 2019, Siegel et al., 2022). It is interesting to see that we found more increasing effects for the crude differences, which disappear when the regression slopes were investigated – indicating that the phenomena of hugely inflated initial effect sizes ("early extreme" or "winner's curse") might not be as pronounced in clinical child psychology as in other fields. Given the fact that we found the decline effects being two times higher in size than the increase effects, we would assume that the underlying reason for the effect misestimation is indeed a genuine decline effect, but without the "winner's curse"-issue of hugely inflated initial effect sizes.

Our average overall summary effect was small (r = .205, p < .001), so we can assume that most of the effects in clinical child psychology are rather small. If one wished to detect a small correlational effect with an alpha probability of 5% (two-tailed) and statistical power of at least 80% (which would be desirable), one would need a sample size of at least 700 *N* (Faul et al., 2009). The median sample size in our sample was *Md* = 111 and about 60% of studies had sample sizes smaller than 200, meaning that most meta-analyses (except the cohort studies) actually had too small sample sizes to detect a small effect. This finding is reflected in the average study power as well, which is just about as high as random chance.

As heterogeneity also might play a role in temporal effect changes, the effects of sample number, total *N*, and average *N* on I^2 and tau^2 was investigated. Total *N* was revealed as a significant positive predictor for both I^2 and tau^2 and average *N* as a positive predictor for I^2. The number of *N* seems to be playing a role in heterogeneity of studies. This is in line with the literature for clinical trials, where we would assume a larger sample to be less heterogeneous (IntHout et al., 2015).

Various factors were considered to find out what the drivers behind these temporal effect changes were. One of the drivers may be the study design. We investigated whether different study designs (i.e., whether studies that are based on means vs studies that are based on correlations) would have an influence on decline effects and increase effects. For the studies based on means we found a small decline and for the studies based on correlations we found a small increase effect. Subgroup analysis showed that these two groups differed statistically significantly from each other. The reason we found differences between the two groups here might be because they were differently affected from publication bias mechanisms. Studies that investigate means usually compare two (or more) groups with each other. Therefore, they need to split their sample size into the groups, meaning there are fewer

participants in each group. Contrary to a correlational design, where only one group is investigated. This has consequences for study power and increases the probability for statistical flukes. We found declines in the studies that investigate means and not in correlational studies, and significant differences between the two groups.

Multiple predictors were tested via multiple hierarchical regression and single regressions to see if certain study characteristics would influence effect misestimation.

Initial study characteristics

The absolute initial effect size revealed itself to be a significant predictor for the crude differences. These findings replicated also with the single regression but only for the complete dataset. In line with the decline effect theory (Protzko & Schooler, 2017), we would assume that larger initial effects would also lead to more decline.

The initial *N* revealed itself as a significant predictor for the regression slopes in the complete and in the published dataset. Single regressions confirmed this picture for both the complete and the published datasets. As with rising *N* in initial studies the statistical power increases, it seems reasonable that the initial *N* has an influence on the decline effect and on the absolute initial effect size. The current impact factor of the initial journal and the annual citations per year of the initial paper and the publication year of the initial study did not have any explanatory value.

Meta-analytic characteristics

The absolute meta-analytic summary effect was a relevant predictor for both the crude differences and the regression slopes in the complete dataset and the published dataset. For single regressions it was only a relevant predictor for the crude differences in the published dataset. The association was consistently negative, indicating that regardless of the underlying data, the higher the summary effect, the smaller the decline. This might be as a higher summary effect might be closer to the high initial effects and therefore the difference between these two is smaller and thereby artificially decreasing the decline effect.

The sample number of the meta-analyses was a relevant predictor in the crude differences only in the complete dataset, but in the regression slopes it displayed significance in both datasets. In single regressions, the sample number was only in the regression slopes (complete and published dataset) a significant predictor. The sample number showed a small negative association. As the regression slopes have more datapoints than the crude differences, it is not surprising that the sample number plays a bigger role here. The negative

association might be a similar one to the numbers of *N*. With increasing datapoints the precision of a study increases. In this case the sample size played a minor role in study power and effect misestimation compared to the actual *N*.

The covered time span was a predictor included in the models for the crude differences and the regression slopes in the complete dataset and in the published dataset for the regression slopes. Although included in the models, it was never a significant predictor, not even in the single regressions. This might be as one of our inclusion criteria was that the meta-analysis mustn't be artificially restricted in time frame anyways and therefore the covered time span did not seem to have any further influence here.

The predictor "unpublished studies" (i.e., whether grey literature was included in a meta-analysis or not) was a significant predictor for the crude differences but not for the regression slopes. Single regressions confirmed this picture, where it was again a significant predictor for the crude differences but not for the regression slopes. The association was positive and small. It would be expected that the inclusion of grey literature had an effect on all the factors above (summary effect, sample numbers, timespan). However, it displayed significance only in the crude differences, which can be seen in the calculations of decline and increase effects above (Table 2). The crude differences also showed a more heterogenous picture regarding declines and increases than the regression slopes, and the increases seemed to stem mostly from published studies.

It is noteworthy that the inclusion of grey literature – which is often named as a strategy to counteract effect misestimation and publication bias – did not seem to make any difference for the findings in this study. All χ^2 tests and Fishers exact tests were non-significant, thereby indicating no influence of grey literature on publication bias. One explanation for this surprising and counterintuitive finding might be that the amount of grey literature in the studies investigated was not sufficient for the tests to detect an influence, because grey literature was not always included and if grey literature was included, it usually made out a few data points. Some publication bias detection methods, as stated in the Methods section, struggle with heterogeneity, and perform worse when heterogeneity is present. Given that sample numbers influenced heterogeneity in our sample, this might be reflected in the publication bias tests.

Publication year was revealed as significant predictor for the regression slopes in the complete dataset as well as in the published one. Single regressions confirmed this finding. The association was positive. This might reflect the fact that with the progress of time, as

studies got larger sample sizes, new methods emerged, and this is revealed in subsequent studies.

One of our main predictions for a possible cause of temporal effect change was publication bias. The occurrence of publication bias was investigated using nine different publication bias detection methods. For almost all methods, indication of bias could be found, with twice as many bias indications when only published literature was investigated. This is strong evidence that publication bias might indeed be the main driver for effect misestimation. If that is the case, future research should consider how to minimise publication bias in the future.

To summarise, first we found evidence for a genuine decline effect but without the "winner's curse"-phenomenon in clinical child psychology. What we also need to address in this context is the usage of diagnostics in clinical empirical research. Unlike other psychological disciplines, the clinical sector works mostly with diagnoses. We know that the changes of diagnostic criteria (e.g., from the *DSM-IV* to the *DSM-5*) influenced clinical samples as people might be considered part of the clinical experimental group when they did not before and vice versa. Of course, this has an influence on the outcomes of our primary studies and in turn on our meta-analytic evidence. Evidence of this change was already found in autism studies (Rødgaard et al., 2019). Thus, this phenomenon can be seen as a genuine driver for effect misestimation. Moreover, it might be the reason for the group difference in studies with means vs. studies with correlations. An RCT or a clinical design is usually based on a group comparison for a novel treatment. The criterion that determines which participant lands in the experimental group and who is part of the control group gets blurred if diagnostic criteria become broader. Changes of diagnostic criteria will influence group assignments and this influence might be diminished when correlational studies are observed, as they investigate only co-occurrences of phenomena.

Second, we found substantial evidence for publication bias since every single publication bias detection method indicated the occurrence of bias (regardless of the presence of grey literature or not). Furthermore, the most influential predictors for the multiple hierarchical and single regressions were all predictors that were related to the sample size of the study. As studies with smaller samples have less power, the probability for statistical flukes increases. Given that the average power we detected across the studies was at chance-level, this is a mechanism we cannot rule out. However, we have to add that we could not find evidence that the citation numbers or Impact Factor had any influence on our effect misestimations.

Regarding the use of publication bias detection methods, we found that they were indeed frequently used, but most methods were older or outdated. Subsequent research on publication bias detection methods found substantial flaws in older publication bias methods, e.g., the Failsafe *N* and the funnel plot were heavily criticised for their unreliance (Sterne et al., 2005). Still, the funnel plot was the most often used tool to investigate publication bias, followed by other small-study-methods as "Trim & Fill" and Egger's regression, and the Failsafe *N*. In eight of 45 cases in our sample, researchers reported the presence of publication bias in their study, in 25 cases they reported that they did not find evidence for publication bias.

For future research, we would recommend to use more modern publication bias detection methods (e.g., the Selection models) and to ensure that the methods used stem from different "publication bias groups" according to the triangulation principle (Kepes et al., 2022), so that the methods can compensate for their respective weaknesses.

Fortunately, the awareness for publication bias has grown among the scientific community in recent years. This has been shown by the fact that journals begin to dedicate special issues to null results, publication bias and file drawer effects (for example in the *JCPP* see Oznoff, 2011, and Oldehinkel, 2018).

Our next question which arose was that, if there is awareness for publication bias among researchers of clinical child psychology, do we find awareness for temporal effect changes as well. To answer this question, we conducted a researcher survey. Asking researchers about their scientific field can give us insights about the intuition and perception of certain phenomena, for example see Fiedler and Schwarz (2016) or Fanelli (2009) for investigating QRPs with researcher surveys or Protzko (2020) for a researcher survey regarding temporal effect changes.

In our survey, the majority of researchers did indeed predict effect sizes becoming smaller over time, which would overlap with the theory of decline effect. Additionally, researchers were indeed aware of the phenomenon of publication bias, since "publication bias" was named as the second most frequent cause for changing effect sizes in the open question. The idea of declining effects also spans into the prediction of the size of the effect. On average, the initial effect size was predicted bigger as the subsequent current effect size.

Researchers predicted medium to large effects for both the initial and the current overall effect. However, when comparing the predicted current effect with the calculated overall summary effect, the estimated effect was a medium effect, whereas the calculated summary

effect was a small one. This disparity between how scientists perceive effect sizes intuitively and how large they really are might mirror the inflated effect sizes we found – that people are "used to" medium and large effects in clinical child psychology and thereby expecting them to be of a certain size. But when investigating average effect sizes in psychology, they tend to be rather small. For example, Gignac and Szodorai (2016) empirically investigated the value of Cohen's (1988) recommendations for effect size interpretation in individual differences research and recommended new interpretation thresholds of .10, .20 and .30 for Pearson *r*, which are much lower than the ones proposed by Cohen.

The frequent answer of "larger samples" as a cause for declining effects would fit into the explanatory approach of decline effect. Due to the frequency of this answer, we also investigated if sample sizes really increased over time. Spearman's rank correlation was conducted and showed a significant positive association between year and sample size (*rho* = .211, $p < .001$), indicating that if a study was younger, this was linked to larger sample sizes. Interestingly, this positive trend was indeed predicted correctly. This is in line with the claims stating that decline effect occurs, because follow-up studies and replications usually have higher statistical demands, which means, among other things, larger sample sizes, higher power, and subsequent smaller effect sizes (Ioannidis, 2005).
Nevertheless, it is important to point out that not all researchers agreed on that proposition of declining effects. More than half of the researchers predicted smaller effects (66%), but a quarter of researchers predicted a different outcome, namely stable (16%) or even increasing (14%) effects. Additionally, publication bias was not the most frequent answer to effect changes, but the improvement of studies and better study designs, which would also partly overlap with the "larger samples" explanation. But clinical explanations, e.g., changes in diagnostic criteria, were given as well – a circumstance that has been proven to influence clinical effect sizes too (for example see Rødgaard et al., 2019). Our approach to consider researchers who have published in the *JCPP* experts was confirmed. Over 60% stated that they would indeed consider themselves experts for clinical child psychology.

To conclude, we confirm that publication bias and decline effect are two phenomena of which the scientific community of clinical child psychology seems to be aware. However, we also find disagreement in regard of effect size direction and the reasons for it. Researchers predicted rather optimistic effects for both the initial and current effect size and used primarily simple and outdated measures to investigate publication bias.

5. Conclusion

The phenomenon of effect misestimation, meaning declining and increasing effects, is present in clinical child psychology as well as publication bias.

Future research should be aware of temporal effect changes and take precautions against it. On the one hand to conduct well designed, adequately powered, replicable studies to improve primary studies in clinical child research in general. Additionally, it is important to be aware of the specifics in the clinical field, e.g., the reliability on diagnostic criteria that change over the course of time. This influences our clinical studies and further our meta-analytic analyses of the field. Future research should ensure to use state of the art diagnostics and be aware that, when looking at older studies, the effect sizes might be inflated.

On the other hand, we should acknowledge that clinical child psychology is not free from publication bias. It is important to stress that publication bias is a result of the need for novel, significant findings as they are more publishable in journals. The publication performance of a researcher is directly linked to one's future career options (Fanelli et al., 2017) and might prompt people to engage into QRPs.

Short-term approaches to improve science in this case include Open Science practices (open access papers, open data, open code, preregistration, etc.), registered reports with provisionally publication guarantee or preprints and testing for publication bias with adequate statistical methods. However, for the future we also need more long-term efforts and structural changes regarding academia – to make publishing findings easier regardless of the outcome and a structural change in the academic system that includes less competition and academic pressure. An example could be changes in journal policies and not to link job options solely on publication performance. All these practices should minimise QRPs and ultimately publication bias in the future.

6. Reference List

2021 Journal Impact Factor, *Journal Citation Reports (Clarivate, 2021)*

Aldhous, P. (2011, May 5). Journal rejects studies contradicting precognition. Retrieved from https://www.newscientist.com/article/dn20447-journal-rejects-studies-contradicting-precognition/#ixzz7EeuzO7rE

Alinaghi, N., & Reed, W. R. (2018). Meta-analysis and publication bias: How well does the FAT-PET-PEESE procedure work? *Research Synthesis Methods*, *9*(2), 285–311. https://doi.org/10.1002/jrsm.1298

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Baker, D., Eslinger, P. J., Benavides, M., Peters, E., Dieckmann, N. F., & Leon, J. S. (2015). The cognitive impact of the education revolution: A possible cause of the Flynn Effect on population IQ. *Intelligence*, *49*, 144–158. https://doi.org/10.1016/j.intell.2015.01.003

Bakker, M., Van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Banks, G. C., & McDaniel, M. A. (2011). The kryptonite of Evidence-Based I–O psychology. *Industrial and Organizational Psychology*, *4*(1), 40–44. https://doi.org/10.1111/j.1754-9434.2010.01292.x

Becker, B. J. (2005). Failsafe N or file-drawer number. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 111–125). John Wiley & Sons, Ltd.

Begg, C. B., & Berlin, J. A. (1988). Publication bias: A problem in interpreting medical data. *Journal of the Royal Statistical Society*, *151*(3), 419. https://doi.org/10.2307/2982993

Begg, C. B., & Berlin, J. A. (1989). Publication bias and dissemination of clinical research. *Journal of the National Cancer Institute*, *81*(2), 107–115. https://doi.org/10.1093/jnci/81.2.107

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*(4), 1088. https://doi.org/10.2307/2533446

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. https://doi.org/10.1037/a0021524

Berger, M. F., & Hersov, L. (2009). JCPP - the journal of child psychology and psychiatry: A history from the inside. *Journal of Child Psychology and Psychiatry*, *50*(1–2), 2–8. https://doi.org/10.1111/j.1469-7610.2008.02036.x

Blease, C., Colagiuri, B., & Locher, C. (2023). Replication crisis and placebo studies: rebooting the bioethical debate. *Journal of Medical Ethics*, jme-108672. https://doi.org/10.1136/jme-2022-108672

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470743386

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

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144. https://doi.org/10.1177/2515245919847196

Clements, J. C., Sundin, J., Clark, T., & Jutfelt, F. (2022). Meta-analysis reveals an extreme "decline effect" in the impacts of ocean acidification on fish behavior. *PLOS Biology*, *20*(2), e3001511. https://doi.org/10.1371/journal.pbio.3001511

Cockburn, A., Dragicevic, P., Besançon, L., & Gutwin, C. (2020). Threats of a replication crisis in empirical computer science. *Communications of the ACM*, *63*(8), 70–79. https://doi.org/10.1145/3360311

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. 2nd (Ed.) Hillsdale, New Jersey: Erlbaum.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2019). *The handbook of research synthesis and meta-analysis* (3rd ed.). Russell Sage Foundation

Cuijpers, P., Smit, F., Bohlmeijer, E. T., Hollon, S. D., & Andersson, G. (2010). Efficacy of cognitive–behavioural therapy and other psychological treatments for adult depression: meta-analytic study of publication bias. *British Journal of Psychiatry*, *196*(3), 173–178. https://doi.org/10.1192/bjp.bp.109.066001

DePalma J. A. (2000). Evidence-based clinical practice guidelines. *Seminars in perioperative nursing*, *9*(3), 115–120.

Duval, S. (2005). The trim and fill method. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), Publication bias in meta-analysis: *Prevention, assessment and adjustments* (pp. 127–144). John Wiley & Sons, Ltd.

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

Duval, S., & Tweedie, R. L. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in Meta-Analysis. *Biometrics*, *56*(2), 455–463. https://doi.org/10.1111/j.0006-341x.2000.00455.x

Easterbrook, P., Gopalan, R., Berlin, J. A., & Matthews, D. R. (1991). Publication bias in clinical research. *The Lancet*, *337*(8746), 867–872. https://doi.org/10.1016/0140-6736(91)90201-y

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. https://doi.org/10.1136/bmj.315.7109.629

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PloS one, 4(5), e5738.*

Fanelli, D., Costas, R., & Ioannidis, J. P. A. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(14), 3714–3719. https://doi.org/10.1073/pnas.1618569114

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41, 1149-1160*

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science, 7(1), 45-52.*

Field, J. B., Bosco, F. A., & Kepes, S. (2020). How robust is our cumulative knowledge on turnover? *Journal of Business and Psychology*, *36*(3), 349–365. https://doi.org/10.1007/s10869-020-09687-3

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, *101*(2), 171–191. https://doi.org/10.1037/0033-2909.101.2.171

Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, *19*(2), 151–156. https://doi.org/10.3758/s13423-012-0227-9

Franco, A., Malhotra, N. R., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

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

Goldacre, B., Drysdale, H., Powell-Smith, A., Dale, A., Milosevic, I., Slade, E., ... & Heneghan, C. (2016). *The COMPare trials project*. Retrieved from COMPare-trials.org.

Gong, Z., & Jiao, X. (2019). Are effect sizes in emotional intelligence field declining? A Meta-Meta analysis. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01655

IntHout, J., Ioannidis, J. P. A., Borm, G. F., & Goeman, J. J. (2015). Small studies are more heterogeneous than large ones: a meta-meta-analysis. *Journal of Clinical Epidemiology*, *68*(8), 860–869. https://doi.org/10.1016/j.jclinepi.2015.03.017

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

Ioannidis, J. P. A., Stanley, T. D., & Doucouliagos, H. (2017). The power of bias in economics research. *The Economic Journal*, *127*(605), F236–F265. https://doi.org/10.1111/ecoj.12461

Ioannidis, J. P. A., & Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology*, *58*(6), 543–549. https://doi.org/10.1016/j.jclinepi.2004.10.019

Ioannidis, J. P. A., & Trikalinos, T. A. (2007a). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*(3), 245–253. https://doi.org/10.1177/1740774507079441

Ioannidis, J. P. A., & Trikalinos, T. A. (2007b). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*(3), 245–253. https://doi.org/10.1177/1740774507079441

Jeschke, J. M., Aparicio, L. G., Haider, S., Heger, T., Lortie, C. J., Pyšek, P., & Strayer, D. L. (2012). Support for major hypotheses in invasion biology is uneven and declining. *NeoBiota*, *14*, 1–20. https://doi.org/10.3897/neobiota.14.3435

Kepes, S., Wang, W., & Cortina, J. M. (2022). Assessing publication bias: A 7-Step user's guide with Best-Practice recommendations. *Journal of Business and Psychology*. https://doi.org/10.1007/s10869-022-09840-0

Light, R. J., Richard, J., Pillemer, D. B., & Light, R. (1984). *Summing up: The science of reviewing research*. Harvard University Press.

Lynn, R. (2009). What has caused the flynn effect? Secular increases in the development quotients of infants. *Intelligence*, *37*(1), 16–24. https://doi.org/10.1016/j.intell.2008.07.008

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below. 05. *The Quarterly Journal of Experimental Psychology*, *65*(11), 2271-2279.

McLeod, B. D., & Weisz, J. R. (2004). Using dissertations to examine potential bias in child and adolescent clinical trials. *Journal of Consulting and Clinical Psychology*, *72*(2), 235–251. https://doi.org/10.1037/0022-006x.72.2.235

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

Neisser, U. (1997). Rising scores on intelligence tests: Test scores are certainly going up all over the world, but whether intelligence itself has risen remains controversial. *American Scientist*, *85*(5), 440-447.

Neuroskeptic. (2012). The nine circles of scientific hell. *Perspectives on Psychological Science*, *7*(6), 643–644. https://doi.org/10.1177/1745691612459519

Nuijten, M. B., Van Assen, M. A., Augusteijn, H. E. M., Crompvoets, E. a. V., & Wicherts, J. M. (2020). Effect sizes, power, and biases in intelligence research: A Meta-Meta-Analysis. *Journal of Intelligence*, *8*(4), 36. https://doi.org/10.3390/jintelligence8040036

Oldehinkel, A. (2018). Sweet nothings–the value of negative findings for scientific progress. *Journal of Child Psychology and Psychiatry*, *59(8)*, *829-830.*

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

Ozonoff, S. (2011). The first cut is the deepest: why do the reported effects of treatments decline over trials?. *Journal of Child Psychology and Psychiatry, 52(7), 729-730.*

Peters, J., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology*, *61*(10), 991–996. https://doi.org/10.1016/j.jclinepi.2007.11.010

Pfeiffer, T., Bertram, L., & Ioannidis, J. P. A. (2011). Quantifying selective reporting and the proteus phenomenon for multiple datasets with similar bias. *PLOS ONE*, *6*(3), e18362. https://doi.org/10.1371/journal.pone.0018362

Pietschnig, J., & Gittler, G. (2015). A reversal of the flynn effect for spatial perception in german-speaking countries: Evidence from a cross-temporal IRT-based meta-analysis (1977–2014). *Intelligence*, *53*, 145–153. https://doi.org/10.1016/j.intell.2015.10.004

Pietschnig, J., & Voracek, M. (2015). One century of global IQ gains: A formal meta-analysis of the Flynn effect (1909–2013). *Perspectives on Psychological Science, 10(3), 282-306.*

Pietschnig, J., Siegel, M., Eder, J. S. N., & Gittler, G. (2019). Effect declines are systematic, strong, and ubiquitous: A Meta-Meta-Analysis of the decline effect in intelligence research. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.02874

Protzko, J. (2020). Kids These Days! Increasing delay of gratification ability over the past 50 years in children. *Intelligence, 80, 101451.*

Protzko, J., & Schooler, J. W. (2017). Decline effects: Types, mechanisms, and personal reflections. *Psychological science under scrutiny: Recent challenges and proposed solutions, 85-107.*

Renkewitz, F., & Keiner, M. (2019). How to detect publication bias in psychological research. *Zeitschrift Für Psychologie*, *227*(4), 261–279. https://doi.org/10.1027/2151-2604/a000386

Rødgaard, E. M., Jensen, K., Vergnes, J. N., Soulières, I., & Mottron, L. (2019). Temporal changes in effect sizes of studies comparing individuals with and without autism: a meta-analysis. *JAMA psychiatry*, 76(11), 1124-1132.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Schooler, J. W. (2011). Unpublished results hide the decline effect. *Nature*, *470*(7335), 437. https://doi.org/10.1038/470437a

Siegel, M., Eder, J. S. N., Wicherts, J. M., & Pietschnig, J. (2022). Times are changing, bias isn't: A meta-meta-analysis on publication bias detection practices, prevalence rates, and predictors in industrial/organizational psychology. *Journal of Applied Psychology*, *107*(11), 2013–2039. https://doi.org/10.1037/apl0000991

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

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). 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*, *144*(6), 1146–1152. https://doi.org/10.1037/xge0000104

Siontis, K. C., Patsopoulos, N. A., & Ioannidis, J. P. A. (2010). Replication of past candidate loci for common diseases and phenotypes in 100 genome-wide association studies. *European Journal of Human Genetics*, *18*(7), 832–837. https://doi.org/10.1038/ejhg.2010.26

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60–78. https://doi.org/10.1002/jrsm.1095

Sterne, J. A. C., Becker, B. J., & Egger, M. (2005). The funnel plot. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 76–98). John Wiley & Sons, Ltd.

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

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams Jr, R. M. (1949). *The american soldier: Adjustment during army life. (Studies in social psychology in World War II)*, Vol. 1. Princeton University Press.

Terrin, N., Schmid, C. H., & Lau, J. (2005). In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *Journal of Clinical Epidemiology*, *58*(9), 894–901. https://doi.org/10.1016/j.jclinepi.2005.01.006

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine*, *22*(13), 2113–2126. https://doi.org/10.1002/sim.1461

Van Aert, R. C. M., Wicherts, J. M., & Van Assen, M. A. (2019). Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis. *PLOS ONE*, *14*(4), e0215052. https://doi.org/10.1371/journal.pone.0215052

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

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

Vankov, I., Bowers, J., & Munafò, M. R. (2014). Article commentary: on the persistence of low power in psychological science. *Quarterly Journal of Experimental Psychology, 67*(5), 1037-1040.

Vevea, J. L., Coburn, K. M., & Sutton, A. J. (2019). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 383–429). Russell Sage Foundation

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods*, *10*(4), 428–443. https://doi.org/10.1037/1082-989x.10.4.428

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*(3), 291-298.

World Health Organization. (2022). *ICD-11: International classification of diseases (11th revision).* *https://icd.who.int/*