Digital mental health interventions for anxiety: An umbrella review and multiverse meta-analysis

Constantin Yves Plessen, MSc[a,b], Olga Panagiotopoulou, BSc[b], Clara Miguel, MSc[b,c], Marketa Ciharova, MSc[b,c], Davide Papola, PhD[b], Darin Pauley, MSc[b], Eirini Karyotaki, PhD[b,c], Pim Cuijpers, PhD[b,c]

[a]Department of Psychosomatic Medicine, Center of Internal Medicine and Dermatology, Charité - Universitätsmedizin Berlin, Berlin, Germany

[b]Department of Clinical, Neuro-, and Developmental Psychology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[c]Amsterdam Public Health Research Institute, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Correspondence to:

Constantin Yves Plessen

Department of Clinical Psychology

Vrije Universiteit Amsterdam

Van der Boechorststraat 7

1081 BT Amsterdam, the Netherlands

Tel: +31 20 598 72 82

Abstract

Several meta-analyses and meta-reviews have yielded inconclusive results on the efficacy of digital mental health interventions for anxiety disorders. To address this discrepancy, we conducted an umbrella review and multiverse meta-analysis of randomized controlled trials with no restrictions on intervention type, target group, format, control condition, or diagnosis. Studies focusing on comorbid anxiety were excluded. We screened 10.942 records and retrieved 1.263 full texts, identifying 10 meta-analyses and 81 randomized controlled trials meeting inclusion criteria. Unfortunately, the majority of included meta-analyses and primary investigations were of suboptimal quality, with possible conflicts of interest. Only two meta-analyses provided suggestive evidence, and the majority yielded weak evidence. Digital interventions demonstrated stronger evidence for the treatment of generalized anxiety disorder than other anxiety disorders. The multiverse meta-analysis calculated 2220 additional meta-analyses, with an average effect size of Hedges' g = 0.65 with 79% of the meta-analyses reaching at least a small effect size. Results suggest the overall robustness of digital interventions, with larger effect sizes observed for guided interventions, internet-based interventions, and comparisons with waitlist control groups. Nonetheless, we caution against overinterpreting findings given the overall low quality of evidence.

1. Introduction

Over the last decade, several meta-analyses and meta-reviews attempted to synthesize the research on the efficacy of digital interventions for mental disorders including anxiety disorders—unfortunately with diverging conclusions. For example, the most comprehensive meta-review synthesizing the results of seven meta-analyses concluded that "apps for anxiety and depression hold great promise with clear clinical advantages, either as stand-alone self-management or as adjunctive treatments" (Lecomte et al., 2020, p.10). The most recent meta-review, however, concluded, that "these results suggest that mobile phone-based interventions may hold promise for modestly reducing common psychological symptoms (e.g., depression, anxiety), although effect sizes are generally small and rarely do these interventions outperform other interventions intended to be therapeutic (i.e., specific active controls)" (Goldberg et al., 2022, p. 15). Both reviews were narrative in nature and did not quantify the overall efficacy of such interventions. Thus, the reasons for diverging conclusions and the robustness of the published evidence remain unclear. Discrepancies like these are harmful as they cause disagreements among policymakers, clinicians and researchers over resource allocation and treatment recommendations (Solmi et al., 2018). Digital interventions—if effective—hold great potential in addressing the supply difficulties faced by individuals seeking mental health services. Such individuals often experience lengthy waiting periods for appointments, limited access to specialized care, and encounter difficulties in accessing mental health services (Chisholm et al., 2016; Patel et al., 2018).

As of January 2023, more meta-analyses have been published on the same overarching research question (i.e. whether digital interventions are effective in treating mental health disorders) with, again, diverging recommendations for anxiety disorders. Two meta-analyses

recommend the usage of apps and online therapy for the treatment of anxiety symptoms (Linardon et al., 2019; Pauley et al., 2021), while another did not (Weisel et al., 2019). Can these differences be understood in light of the fact that all stages of conducting a meta-analysis consist of countless small judgments and decisions (e.g., deciding which studies to include or exclude, how to handle outliers, which statistical method to use for pooling the studies, and how to interpret the findings of a meta-analysis) (Sharpe & Poets, 2020)? Or do substantial differences explain these diverging results, for instance, the type of anxiety disorder, or the technology used to deliver the intervention?

Thus, the research field on digital psychological interventions for anxiety appears fragmented, because these meta-analyses focus on different interventions (internet-based, smartphone-based, guided or self-guided), populations (clinical or community populations), anxiety disorders (generalized, social anxiety disorder, panic disorder with or without agoraphobia, mixed anxiety disorders, other anxiety disorders like specific phobias) and different methodologies are used to synthesize the evidence. In addition, the impact of the poor quality of meta-analyses and primary studies in this field on the reliability of results is uncertain. As a result, a thorough examination of the entire field is necessary. To fill this gap, we present the results of two methods for research synthesis: 1) a systematic umbrella review and 2) a multiverse meta-analysis.

An umbrella review (or meta-review) is a systematic review of all meta-analyses of the existing literature, allowing for a higher-level synthesis of the data and a better recognition of uncertainties, biases, and knowledge gaps than conventional meta-analyses (Sharpe & Poets, 2020). Within this framework, we investigated the quality, the unique number of primary trials, the validity of their designs, the various types and foci of therapies, and the evidence for the

efficacy of treatment in each meta-analysis. This method enabled us to investigate the current meta-analytic evidence on the efficacy of digital mental health interventions for anxiety disorders.

In our multiverse meta-analysis, we went a step further and investigated the robustness of the meta-analytic evidence by conducting all thinkable meta-analyses using data from randomized controlled trials on patients with elevated anxiety symptoms or a clinical diagnosis of any anxiety disorder. These studies were either already part of the meta-analyses included in our umbrella review or were identified through a targeted search. Such a multiverse meta-analysis conducts all possible meta-analyses simultaneously and enables us to visualize the emerging results of thousands of meta-analyses at once (Plessen et al., 2023; Voracek et al., 2019). As a result, we were able to investigate whether most meta-analyses that could be conducted with this body of literature would produce clinically relevant effect sizes.

With this combination of methods, we aim to give a comprehensive overview of the evidence that is contained in all existing meta-analyses in the field of digital interventions for anxiety disorders.

2. Methods

2.1. Data

2.1.1. Identification and Selection of Studies

We systematically searched for meta-analyses of randomized controlled trials and randomized controlled trials on the effectiveness of digital interventions for anxiety. A targeted search was performed on February 6th 2023 as part of a larger project that is to create a database with all available studies (randomized controlled trials, observational studies, systematic reviews and meta-analyses) that focus on digital interventions for the treatment of anxiety and/or depression. For a full description of the search strings, please see Appendix A. We followed general guidelines for conducting and reporting umbrella reviews (Ioannidis, 2009, 2017; Papatheodorou, 2019; Solmi et al., 2018). We registered our protocol for this umbrella review at the Open Science Framework (Plessen, 2021; https://osf.io/dm9x3/).

We included meta-analyses and randomized controlled trials following this PICO process:

- Patients: Adults above the age of 18 diagnosed with an anxiety disorder or scoring above a cut-off on a self-reported instrument assessing anxiety severity. We also included patients with elevated yet subclinical symptoms.
- Intervention: Digital interventions as a standalone treatment. This refers to any psychological intervention delivered by any electronic device (e.g., a smartphone app, an internet browser of a phone or computer) and can either be guided, or self-guided.

- Comparison: We considered studies that compare an intervention group with any control condition, either passive (care as usual, waitlist), or active (psychoeducation, placebo, attention control, but not another psychological intervention, like face-to-face therapy or another digital intervention).
- Outcomes: Any instrument assessing anxiety: generalized anxiety disorder, social anxiety disorder, panic disorder, and specific phobias.

We only included meta-analyses and randomized controlled trials that reported effect sizes or at least enough information to calculate standardized mean differences post-intervention (Hedges' *g*). We included meta-analyses that included controlled but non-randomized studies if the results were presented separately for randomized and non-randomized trials. Meta-analyses of trials that were not only focused on anxiety (for instance, meta-analyses focusing on mental disorders in general) were also included, but only if effect sizes for addressing anxiety were reported separately. We only included meta-analyses published after 2017, as we expected a large amount of small meta-analyses with a large overlap between them.

We systematically screened three databases: Pubmed (N = 7.302), Embase (N = 8.729), and PsycINFO (N = 3.548). Upon removal of the duplicate records (N = 8.749), two independent researchers screened the titles and abstracts of the remaining records (N = 10.830). All records were screened by two independent researchers after duplicates were removed. When at least one of them considered a study would fulfil the criteria for inclusion, the entire text of the study was obtained. Two separate reviewers read the full-text studies, and conflicts were resolved through discussion. For the multiverse meta-analysis, we used both randomized controlled trials included in the meta-analyses that were part of the umbrella review and/or identified from the targeted search.

2.1.2. Data Extraction

For each included meta-analysis, we extracted all reported information on effect sizes (between-groups effect size (Hedges' g) at post-test of each included primary study, 95% confidence interval [CI], level of heterogeneity I^2 with respective 95% CI) (Higgins & Thompson, 2002). We included any instrument (self-rated or clinician-rated) assessing anxiety. If included meta-analyses reported different continuous effect size estimates, such as mean change scores from baseline, for an included primary study, we calculated the post interventions between-group standardized mean difference effect size based on the means, standard deviations and sample sizes given in the primary study.

For each primary study included in the meta-analyses, we first extracted the information from each included meta-analysis. In case insufficient data was provided in these meta-analyses, we extracted the relevant data from the primary study. For all randomized controlled trials that were identified exclusively from the targeted search, data were directly extracted from the published papers. We extracted information on participants (region, recruitment setting), type of digital intervention (guided/self-guided), type of delivery method (internet, smartphone app), type of control group (waitlist/care as usual, active control group), diagnosis (generalized, social, panic, specific phobia, mixed), and risk of bias assessment. We extracted relevant data to conduct the multiverse meta-analysis from the primary studies if this information was missing in the included meta-analyses (e.g., information on moderators, means, standard deviations, sample size). In meta-analyses involving multiple mental disorders, we only included data from trials involving participants with anxiety. The same applied to meta-analyses combining RCT and non-RCT designs, where we only extracted data from RCTs.

Our focus remained on those primary studies conducted on individuals exhibiting elevated symptoms or having a clinical diagnosis of any anxiety disorder. We excluded primary studies undertaken in populations diagnosed with other conditions such as elevated depressive symptoms or elevated stress, or those from the general (healthy) population. We reasoned that assessing treatment efficacy in individuals that are not selected based on pre-existing anxiety levels would introduce excessive noise. Sensitivity analyses were carried out to determine the impact of our strict inclusion criteria (namely randomized controlled trials in people with elevated anxiety symptoms) compared with all studies that were present in the included meta-analyses (thus, including also studies with unselected individuals from the general population).

2.1.3. Data Harmonization Across Meta-Analyses

There were several instances where conflicting data on the same primary study was presented in multiple meta-analyses. Discrepancies were particularly apparent in instances when the same primary study received different risk of bias ratings across multiple meta-analyses. In response, we conducted Cochrane Risk of Bias 1 ratings to determine the risk of bias level for each study—high, moderate, or low—because most of the included primary trials were rated with this tool. We opted for the more conservative rating in ambiguous situations. Such deviations were substantial at times, the same primary study could experience a risk of bias classification ranging from low to high.

Inconsistencies were also observed when multiple effect size estimates reported in meta-analyses could not be reproduced or diverged between meta-analyses. To address this, we calculated these uncertain effect sizes from the raw data provided in each individual trial. We used the extracted means, standard deviations, and sample sizes to compute Hedges g and subsequently replaced these values for all analyses.

In some instances it was not clear which instrument was used in the meta-analyses, we therefore opted to include all relevant additional measures we discovered in the primary studies and incorporated them into our multiverse meta-analyses. At least two raters extracted data for each each meta-analysis and a random subset of 20% of primary studies was also double-coded.

2.1.4. Open Science Practices

We disclosed all data exclusions (if any), all manipulations, and all measures in the study (Simmons et al., 2012). Deviations from the analysis plan were explicitly stated. All components necessary for reproducible data analysis (open data, open code) were made accessible via the OSF and comply with the FAIR (findable, accessible, interoperable, reusable) guiding principles for scientific data (Wilkinson et al., 2016).

2.2.Umbrella Review

2.2.1. AMSTAR 2 Ratings

We used the AMSTAR 2 tool (a critical evaluation tool for systematic reviews) to assess the quality of the included meta-analyses (Shea et al., 2017). AMSTAR-2 evaluates several key aspects of systematic reviews, including (1) the use of the PICO (participants, intervention, comparator, outcome) acronym in formulating the research question, (2) whether the methods were established before the conduct of the review, (3) an explanation for the selection of study design to be included, (4) the comprehensiveness of the search strategy, (5) study selection by at least two reviewers, (6) data extraction by at least two reviewers, (7) providing a list of excluded studies with reasons, (8) a detailed description of included studies, (9) assessment of risk of bias in included studies, (10) the review reported sources of funding for the included studies, (11) appropriate methods for pooling results of individual studies, (12) assessment of the impact of

risk of bias on the outcomes, (13) a discussion of the impact of risk of bias on results, (14) an explanation and discussion of heterogeneity, (15) assessment of publication bias, (16) reporting potential conflict of interest and funding for the review. We assigned positive (yes), probably positive (partial yes), probably negative (partially no), or negative (no) judgements to each item. All of these criteria were assessed by two separate researchers (CM and CYP), and any discrepancies were resolved through conversation or, if necessary, consultation with a third reviewer (PC). We followed the recommendation to provide a global quality rating (e.g., high, moderate, low, critically low). Critical domains were items 2, 4, 9, 11, 13, 15. If no or only one non-critical weakness was detected, we rated the study as high, if more than one non-critical weakness was detected we rated it as moderate, if one critical flaw with or without non-critical weaknesses was detected we rated the study as low, and if more than one critical flaw with or without non-critical weaknesses was detected the study was rated as critically low.

We additionally calculated an AMSTAR 2 score by rating each of the 16 items with a score of 0 (answer “no”), 1 (answer “yes”) or 0.5 (answer “partial yes”) and summing it up. We also converted this score to a percentage (%) summary value.

2.2.2. Proportion of Unique Primary Studies

We collected the references of the primary studies included in each meta-analysis to provide an overview of the proportion of unique studies. We calculated whether each primary study was included in one or more meta-analyses. If a primary study was only included in one meta-analysis, we designated it as a unique primary study for this meta-analysis. The proportion of unique primary studies in a meta-analysis can be interpreted as an indication of the necessity of that meta-analysis, meaning that the greater the number of unique primary studies, the greater the need for this particular meta-analysis.

2.2.3. Quality Assessments for primary studies

For risk of bias assessment, we extracted the quality assessment of the primary studies included in the meta-analyses. When this information was not reported for each primary study, e.g. when only an overall risk of bias score of all included primary studies were reported, two independent researchers assessed the risk of bias based on the Cochrane Risk of Bias Tool 1. As we expected that different instruments were used to assess the quality or risk of bias of the included trials, we transformed the quality score to a proportion by dividing the number of fulfilled criteria by the number of total criteria. The resulting quality score between 0 and 1 allowed us to retrieve comparable ratings across meta-analyses. In addition, we created a categorical variable (0 = low concern; 1 = some concern, 2 = high risk of bias) from these measures according to the standards of the utilized protocol. If multiple meta-analyses reported different values for the same primary studies, we used the most conservative one.

2.2.4. Strength of Evidence

We rated the strength of the evidence for each meta-analysis based on criteria outlined by Fusar-Poli and Radua (2018), namely: 1) the sample size, 2) the *p*-value for the effect size, 3) significant publication bias, and 4) a prediction interval that did not include zero. Based on these criteria, four different classes of evidence could be reached:

- Class I (convincing evidence): number of cases > 1000, *p*-value of the meta-analysis < 10e-6, I^2 < 50%, 95% prediction interval excluding the null, *p*-value of the Egger's test > .05 and *p*-value of the Ioannidis' test > .05

- Class II (highly suggestive evidence): number of cases > 1000, *p*-value of the meta-analysis < 10e-6, largest study with a statistically significant effect and class I criteria not met
- Class III (suggestive evidence): number of cases > 1000, *p*-value of the meta-analysis < 10e-3 and class I–II criteria not met
- Class IV (weak evidence): *p*-value of the meta-analysis < 0.05 and class I–III criteria not met
- Class ns (not statistically significant): *p*-value of the meta-analysis >= 0.05

The *metaumbrella* R package was used for these analyses (Gosling et al., 2023).

2.3.Multiverse Meta-Analysis

At multiple stages in any meta-analysis, researchers have to decide between several equally defensible choices (e.g., different study inclusion criteria, different ways of excluding outliers, different ways of dealing with low quality studies, different choices of effect size estimators, different ways of handling effect size dependency, etc.). In a multiverse meta-analysis, researchers identify all of these possible stages for decisions, determine alternative analysis steps at each stage, and implement all of them. As a result, a multiverse meta-analysis reports the outcomes of all meta-analyses resulting from all possible combinations.

2.3.1. Which Factors

At the stage of study inclusion and exclusion, we included several alternative decisions (so called *Which Factors*, indicating which data to meta-analyze), following the PICOs process.

- For specific patient *populations* we performed meta-analyses focusing on 1) different recruitment strategies (recruiting participants only from the community, from clinical settings, or including all participants).
- Regarding different *interventions*, we conducted meta-analyses focusing on 2) different types of delivery (via internet, smartphone app, or either), and 3) types of guidance (guided, self-guided, or either).
- We also performed meta-analyses with different 4) types of *control* groups (care as usual/waitlist, active, other control groups, or either).
- We conducted meta-analyses focusing on different 5) *outcomes* for anxiety (generalized anxiety disorder, social anxiety disorder, panic disorder, phobias, mixed anxiety, or either diagnosis).
- We also decided to perform different meta-analyses based on *study* type in regard to 6) risk of bias assessment (low risk of bias, some concern). There are several common practices in dealing with the so called "garbage in garbage out" problem which arises when primary studies of poor quality are included in a meta-analysis. In our risk of bias *Which* factor, we include three approaches: a) including all studies, regardless of quality (all risk of bias assessments), b) excluding the high risk of bias studies by following the advice by Eysenck (1995), and c) using a *best evidence synthesis* strategy and including only low risk of bias by following the advice of Slavin (1995).

As a result, our multiverse meta-analysis contains every single meta-analysis that could possibly be conducted with any combination of these subgroups. Consequently, we did not only replicate all meta-analyses that have been conducted and are included in our umbrella review,

but also created a vast amount of new knowledge by conducting hundreds of additional meta-analyses that were still missing in the literature.

2.3.2. How Factors

In addition to these seven *Which Factors*, we also included two *How-Factors* (i.e., decisions about how to analyze the data). Meta-analysts could decide to deal with effect size dependency (i.e. introducing dependent effect sizes by including multiple effect sizes per primary study and thereby violating the assumption of independence, which would be necessary for adequate summary effect size estimation) in several ways: Researchers a) could simply ignore this dependency (we do not believe this to be an "equally defensible analytic strategy", but we decided to include it nonetheless because it is a very prevalent practice and we were interested in its influence on the overall pattern of evidence), b) researchers could model the nested structure by using 3-level models, or c) they could average the effect sizes in each study before estimating the summary effect size.

As a second *How Factor*, we included eight different meta-analytical estimators to analyze the data: a random-effects model, a fixed-effect model, 3-level models, robust-variance-estimation, Weighted Average of Adequately Powered, Unrestricted Weighted Least Squares, PET-PEESE, and *p*-uniform* (an effect size estimate accounting for the potential presence of publication bias). All analyses were performed in R (version 4.2.2). We used predominately the R packages *metafor* and *puniform* for these analyses (van Aert & van Assen, 2020; Viechtbauer, 2010).

We used four methods to address small study effects that can arise due to publication bias, reporting bias, or clinical heterogeneity, which can result in overestimated treatment

effectiveness. PET-PEESE is a regression-based method that corrects for small-study effects in meta-analyses by adjusting for the relationship between effect sizes and standard errors (Bartoš et al., 2022; Stanley, 2017). P-uniform* is a selection model approach that uses a random-effects model to distinguish between statistically significant and non-significant effect sizes and corrects for publication bias by assuming a constant probability of publishing a statistically significant or non-significant effect size (van Aert & van Assen, 2020). Both methods aim to provide more accurate estimates of effect sizes and to correct for potential biases in meta-analyses.

The study also utilized the Unrestricted Weighted Least Squares (UWLS) estimator, which is suggested to have superior statistical properties compared to random-effects models in both the presence and absence of publication bias (Stanley et al., 2022). The advantage of the UWLS estimator is that it can handle correlated heterogeneity and does not assume that the variances of heterogeneity and sampling error are independent. UWLS can provide a more accurate summary effect size, particularly when the variance of heterogeneity is associated with the variance of sampling error. For meta-analyses with studies of varying sample sizes and heterogeneity, UWLS may be particularly useful.

Additionally, the study included a version of the UWLS estimator called the Weighted Average of Adequately Powered (WAAP) estimator (Stanley et al., 2022). WAAP completely excludes small studies and only includes studies with 80% or higher statistical power. When there is publication-selection bias, WAAP is less biased than other weighted average estimators, such as the random-effects, fixed-effect, and UWLS estimators. Simulations suggest that WAAP can significantly reduce bias in the results.

2.3.3. Descriptive Specification Curve

We visualized the emerging evidence of thousands of meta-analyses with a so-called descriptive meta-analytic specification curve plot. This plot allowed us to inspect gaps and patterns in the meta-analytic summary effects for all Which and How Factor combinations sorted by magnitude. We would consider the current evidence for the efficacy of digital mental health interventions for anxiety to be robust if more than 80% of meta-analyses including only high--quality primary studies produced summary effect sizes larger than at least a small effect size of Hedges` $g = 0.20$. We tested this assumption by inspecting the descriptive specification curve plots and reported the percentage of meta-analyses exceeding this threshold. The descriptive meta-analytic specification plot displays the meta-analyses contained in the multiverse of all possible meta-analyses and visualizes each specification's Which and How Factor combination, including the resulting meta-analytic summary effects with their respective 95% confidence intervals.

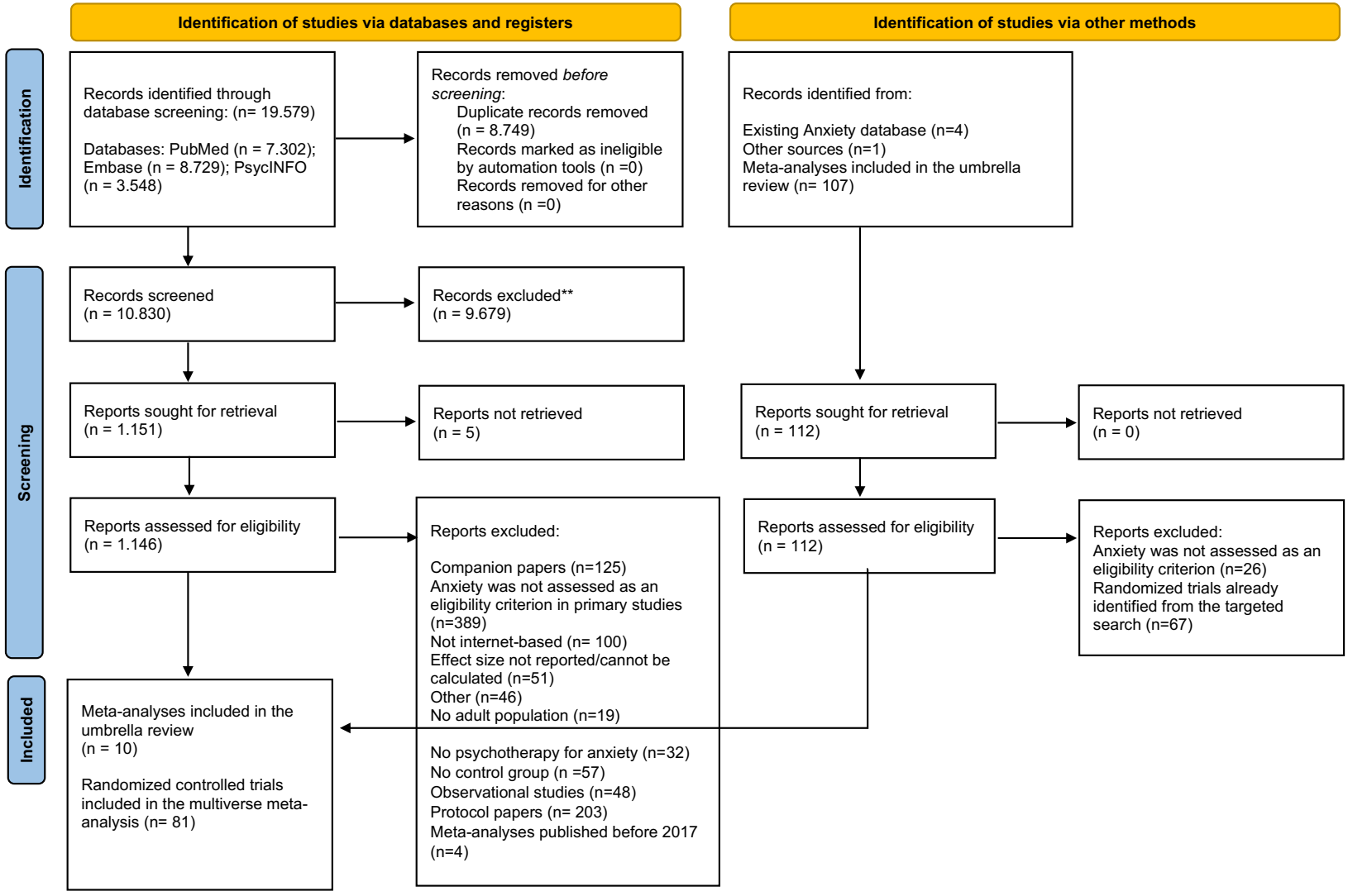
3. Results

3.1.Umbrella Review

3.1.1. Selection and Inclusion of Meta-analyses

In total, the targeted search resulted in 19.579 records, 10.830 after the removal of duplicates. 112 additional records were identified from other sources. We retrieved 1.258 full-text articles and excluded 1.167. Figure 1 shows a PRISMA flowchart describing the inclusion process, with reasons for exclusion. A total of 10 meta-analyses were included in the umbrella review, with 62 unique primary studies, including 188 effect sizes. In the multiverse meta-analysis, we obtained a total of 81 randomized controlled trials. Out of these studies, 63 were selected from the targeted search and 18 were retrieved from other sources.

Figure 1. *PRISMA Flowchart of the Inclusion of Meta-Analyses*


Identification of studies via databases and registers
Identification of studies via other methods
Identification
Screening
Included
Records identified through database screening: (n= 19.579)
Databases: PubMed (n = 7.302); Embase (n = 8.729); PsycINFO (n = 3.548)
Records removed *before screening*:
Duplicate records removed (n = 8.749)
Records marked as ineligible by automation tools (n =0)
Records removed for other reasons (n =0)
Records identified from:
Existing Anxiety database (n=4)
Other sources (n=1)
Meta-analyses included in the umbrella review (n= 107)
Records screened (n = 10.830)
Records excluded** (n = 9.679)
Reports sought for retrieval (n = 1.151)
Reports not retrieved (n = 5)
Reports sought for retrieval (n = 112)
Reports not retrieved (n = 0)
Reports assessed for eligibility (n = 1.146)
Reports excluded:
Companion papers (n=125)
Anxiety was not assessed as an eligibility criterion in primary studies (n=389)
Not internet-based (n= 100)
Effect size not reported/cannot be calculated (n=51)
Other (n=46)
No adult population (n=19)
No psychotherapy for anxiety (n=32)
No control group (n =57)
Observational studies (n=48)
Protocol papers (n= 203)
Meta-analyses published before 2017 (n=4)
Reports assessed for eligibility (n = 112)
Reports excluded:
Anxiety was not assessed as an eligibility criterion (n=26)
Randomized trials already identified from the targeted search (n=67)
Meta-analyses included in the umbrella review (n = 10)
Randomized controlled trials included in the multiverse meta-analysis (n= 81)

3.1.2. Characteristics of the 10 included meta-analyses

Most included meta-analyses focused on specific anxiety disorders, namely generalized anxiety disorder (Eilert et al., 2021), social anxiety disorder (Guo et al., 2021), panic disorder with or without agoraphobia (Domhardt et al., 2020; Stech et al., 2020), or specific phobias (Mor et al., 2021). Two meta-analyses focused on several anxiety disorders in the same review (Pauley et al., 2021; Romijn et al., 2019). One meta-analysis focused on anxiety symptoms without specifying the anxiety disorder (Firth et al., 2017) and two meta-analyses focused on common mental health problems, with a subset of the meta-analysis on anxiety disorders (Linardon et al., 2019; Weisel et al., 2019). Three meta-analyses examined smartphone interventions (Firth et al., 2017; Linardon et al., 2019; Weisel et al., 2019), while the other meta-analyses focused either entirely on internet-based interventions (Eilert et al., 2021; Guo et al., 2021; Romijn et al., 2019; Stech et al., 2020) or included all types of digital interventions (Domhardt et al., 2020; Mor et al., 2021; Pauley et al., 2021). See Tables 1 and 2 for a detailed description of the characteristics, quality ratings and effect sizes of each included meta-analysis. Of the 10 included meta-analyses, only one was aimed at studies on self-guided interventions (Weisel et al., 2019), two meta-analyses primarily included guided interventions (Linardon et al., 2019; Mor et al., 2021), and all other meta-analyses investigated all types of guidance in interventions. Regarding trial designs, the majority of meta-analyses included exclusively randomized controlled trials (RCT), whereas three meta-analyses included non-randomized controlled clinical trials as well (Domhardt et al., 2020; Mor et al., 2021; Stech et al., 2020).

The number of included primary studies in each of the 10 meta-analyses ranged from 3 to 46, the proportion of unique trials included in the meta-analyses ranged from 0 to 100%, while

half of the meta-analyses provided at least 2 unique studies. The total number of included participants per study ranged from $N = 267$ to $N = 4727$ participants.

Table 1. Characteristics and Effect sizes found in meta-analyses of digital interventions for anxiety.

Meta-Analysis	*Focus*	*Participants*	*Intervention*	*Guidance*	*Control*	*Primary Studies*	*Dependency Strategy*	*Estimator*	*Risk of Bias Tool*	*Outcome*	*Sample Size*	*kStudies*	*SMD*	*95% CI*
Domhardt 2020	PD/a	Adults (≥ 18 years), who meet diagnostic criteria for PD/a	IMI	All	WLC	RCT and CCT	Selection	REML	ROB-1	PD	NR (502)	9	NR (1.41)	NR (0.86-1.96)
										Agoraphobia	NR (319)	6	1.15	0.74-1.56
										GAD	NR (290)	5	1.06	0.62-1.50
Eilert 2021	GAD	Adults with clinical diagnosis of GAD who may have had comorbidities and/or impairment in functioning	Internet-Delivered	All	Wait-List, Placebo, Or Attention Control	RCT	Aggregate	REML	CLEAR NPT	GAD	1178	19	0.79	0.55-1.03
										Worry	903	14	0.75	0.53-0.97
Guo 2020	SAD	Adults diagnosed in a clinic or online with mild to moderate SAD	iCBT	All	WLC	RCT	Aggregate	REM	ROB-1	SAD	1200	14	0.79	0.67-0.92
Mor 2021	Specific Phobia	Children, adolescents, or adults who had a diagnosis of sp or presented high scores on self-report measures for phobia	IMI, VR	All	WLC	RCT	Ignored	REML	NHLBI	Specific Phobias	298	3	1.07	0.51-1.62
Pauley 2021	Anxiety Disorders	Adults with clinician validated diagnosis of any primary anxiety disorder.	Digitally Delivered	All	WLC Or CAU	RCT	Selection	DL	ROB-2	GAD, PD/a, Mixed, SAD	4958	47	0.8	0.68–0.93
										GAD	1203	9	0.62	0.31-0.93
										Mixed	958	9	0.68	0.39-0.97
										PD/a	837	15	1.08	0.77-1.39
										SAD	1960	20	0.76	0.62-0.91
Romijn 2019	Anxiety	Adults recruited from the community and through outpatient clinics with a primary diagnosis of an anxiety disorder.	iCBT	All	WLC	RCT	Selection	DL	ROB-1	GAD	2920	41	0.72	0.60-0.83
Stech 2019	PD with or without Agoraphobia	Adults with elevated symptoms of panic disorder (according to validated self-report or clinician-rated scales) and/or a diagnosis of panic disorder	iCBT	All	WLC	RCT and Non-RCTs	Selection	REM	ROB-1, ROBINS-1	PD	NR	9	1.22	0.72-1.71
										Agoraphobia	NR	6	0.91	0.36-1.45
Smartphone-Based Interventions														
Firth 2017	Anxiety	Adults, participation was not restricted by diagnostic status, medication usage, or any other sample characteristics.	Smartphone-Supported	All	All	RCT	Select (Active Comparison)	DL	ROB-1	Mixed	1837	9	0.33	0.173-0.477
Linardon 2019	Common Mental	Adults with elevated anxiety symptoms (9 studies), 19	Smartphone-	All	All	RCT	Ignore	DL	ROB-1	GAD	NR (3867)	28	0.30	0.20-0.40

Table 1. Characteristics and Effect sizes found in meta-analyses of digital interventions for anxiety.

Meta-Analysis	*Focus*	*Participants*	*Intervention*	*Guidance*	*Control*	*Primary Studies*	*Dependency Strategy*	*Estimator*	*Risk of Bias Tool*	*Outcome*	*Sample Size*	*kStudies*	*SMD*	*95% CI*
	Disorders and General Well-Being	studies targeted general well-being	Supported											
										GAD	NR	9	0.36	0.25-0.47
										SAD	NR	6	0.58	0.25-0.90
Weisel 2019	Common Mental Disorders	Adults above clinically relevant cutoffs for anxiety disorders.	Smartphone-Standalone	Unguided	All	RCT	Aggregate	DL	ROB-1	Mixed	479	3	0.3	-0.10-0.70

Note. For meta-analyses that also compared digital interventions with other digital interventions or face-to-face therapies, only information is shown for comparsions with control groups. When information was not reported (NR) we extracted the missing data and report it in brackets after NR; Focus: Focus of included meta-analysis. We included summary effect sizes from meta-analyses reporting effect sizes obtained from primary studies with individuals that were included based on elevated anxiety symptoms. Sometimes these meta-analyses were exclusively on anxiety disorders, sometimes on other disorders as well; IMI: Internet- and mobile-based interventions; RCT: Randomized Controlled Trial, CCT: Controlled Clinical Trial (we only included effect sizes obtained from RCTs); Dependency Strategy: Method used when multiple effect sizes per included primary study were encountered; Estimator: Meta-analytical estimator used (DL = DerSimonian-Laird estimator; REML = restricted maximum likelihood estimator; REM = random effects model, not specified); kStudies: number of studies; NR: Not reported; SMD: standardized mean difference; WLC: Waiting list; RoB1: The Cochrane Collaboration's tool for assessing risk of bias in randomised trials (Higgins et al., 2011); RoB2: Revised Cochrane risk-of-bias tool for randomized trials (Higgins et al., 2020); CLEAR NPT: Checklist to evaluate a report of a nonpharmacological trial (Boutron 2005); NHLBI: The Study Quality Assessment Tools from the National Heart Lung and Blood Institute; ROBINS-1: A tool for assessing risk of bias in non-randomised studies of interventions (Sterne et al., 2016)

Table 2. *Selected characteristics of included meta-analyses*

								AMSTAR[b]																	
Meta-Analysis	***Studies (k)***	***Effect Sizes (k)***	***Unique Studies (k)***	***Unique Studies (%)***	***N[a)]***	***Quality Score (%)***	***Evidence Class***	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***	***15***	***16***	***Rating***	***Total Y***	***Total Y (%)***
Domhardt 2020	9	22	0	0	407	67	IV	Y	Y	N	PY	Y	Y	N	Y	Y	N	N	N	Y	Y	N	Y	Critically low	9.5	59.4
Eilert 2021	17	21	5	29	1,172	67	IV	Y	N	Y	Y	N	N	N	Y	PY	N	Y	N	N	Y	Y	N	Critically low	7.5	46.9
Firth 2017	3	3	0	0	514	60	ns	Y	N	Y	PY	Y	N	N	Y	Y	N	Y	N	N	Y	Y	N	Critically low	8.5	53.1
Guo 2020	14	14	1	7	1,082	67	IV	N	N	Y	PY	N	N	N	N	Y	N	Y	N	PY	Y	Y	Y	Low	7.0	43.8
Linardon 2019	3	7	1	33	352	60	ns	Y	Y	Y	PY	N	N	N	Y	PY	N	Y	Y	Y	Y	Y	N	Moderate	10.0	62.5
Mor 2021	3	6	3	100	267	76	IV	Y	PY	Y	PY	Y	Y	N	PY	Y	N	Y	N	N	Y	N	Y	Critically low	9.5	59.4
Pauley 2021	46	53	8	17	4,727	74	III	Y	N	Y	PY	Y	Y	N	PY	Y	N	Y	Y	Y	Y	Y	Y	Low	12.0	75.0
Romijn 2019	32	41	2	6	1,881	56	I	Y	N	Y	PY	Y	Y	N	Y	Y	N	Y	N	N	Y	Y	Y	Critically low	10.5	65.6
Stech 2019	9	16	0	0	532	57	IV	Y	Y	Y	PY	Y	N	N	Y	Y	N	Y	Y	Y	Y	Y	Y	Moderate	12.5	78.1
Weisel 2019	3	5	0	0	455	40	ns	Y	Y	Y	PY	Y	N	N	Y	Y	N	Y	N	Y	Y	N	Y	Low	10.5	65.6

a) N = sample size of participants included in the meta-analysis fulfilling inclusion criteria of our umbrella review. If the meta-analysis included multiple effect sizes per primary study, we pooled the participants in such instances. Unique studies = we report unique studies that were not part of another meta-analysis. This is based on the primary studies and not whether the exact same effect size was reported, i.e., from the same instrument or the same treatment comparison. Quality score of primary studies (%) = from 0 to 100, the number of fulfilled criteria divided by the number of total criteria for a given risk of bias tool multiplied by 100. Evidence Class of the meta-analysis, according to Ioannidis criteria: Class I (convincing evidence): number of cases > 1000, *p*-value of the meta-analysis < 10e-6, I^2 < 50%, 95% prediction interval excluding the null, *p*-value of the Egger's test > .05 and *p*-value of the Ioannidis test > .05; Class II (highly suggestive evidence): number of cases > 1000, *p*-value of the meta-analysis < 10e-6, largest study with a statistically significant effect and class I criteria not met; Class III (suggestive evidence): number of cases > 1000, *p*-value of the meta-analysis < 10e-3 and class I–II criteria not met; Class IV (weak evidence): *p*-value of the meta-analysis < 0.05 and class I–III criteria not met; Class ns (not statistically significant): *p*-value of the meta-analysis >= 0.05

b) The AMSTAR-2 items refer to 1. the use of the PICO (participants, intervention, comparator, outcome) acronym in formulating the research question, 2. whether the methods were established before the conduct of the review, 3. an explanation for the selection of study design to be included, 4. the comprehensiveness of the search strategy, 5. study selection by at least two reviewers, 6. data extraction by at least two reviewers, 7. providing a list of excluded studies with reasons, 8. a detailed description of included studies, 9. assessment of risk of bias in included studies, 10. the review reported sources of funding for the included studies, 11. appropriate methods for pooling results of individual studies, 12. assessment of the impact of risk of bias on the outcomes, 13. a discussion of the impact of risk of bias on results, 14. an explanation and discussion of heterogeneity, 15. assessment of publication bias, 16. reporting potential conflict of interest and funding for the review. Rating: We followed the recommendation to provide a global quality rating (e.g., high, moderate, low, critically low) based on theoretical considerations of the importance of the domains relevant for our research questions. We consider these critical domains to be items 2, 4, 9, 11, 13, 15. Total Y and Total Y (%): Number of summed positive ratings (PY = 0.5, Y = 1).

All meta-analyses reported some type of overall quality rating for the included studies. The majority of meta-analyses (seven meta-analyses) used the Cochrane Risk of Bias Tool 1 (Higgins et al., 2011), while the other four meta-analyses used either the Cochrane Risk of Bias Tool 2 (Sterne et al., 2019), ROBINS-1 (Risk Of Bias In Non-randomised Studies - of Interventions; Sterne et al., 2016), CLEAR NPT Checklist (Boutron et al., 2005), or Study Quality Assessment Tool of the National Heart, Lung, and Blood Institute NHLBI (NHLBI, 2020). However, only four meta-analyses provided enough information to retrace which study was rated as having a low, high, or medium risk of bias (Mor et al., 2021; Pauley et al., 2021; Stech et al., 2020). The other seven meta-analyses used the risk of bias assessment either for sensitivity analyses (Linardon et al., 2019; Firth et al., 2017), provided a total summary of the risk of bias for all studies (Eilert et al., 2021; Guo et al., 2021) or rated the individual domains of the respective risk of bias tool without providing enough information to calculate a summary score (Domhardt et al., 2020; Weisel et al. ,2019; Romijn et al., 2019). The average standardized quality score (with a score between zero and 1) for all included trials in the meta-analyses, ranged from 0.4 to 0.76.

The pooled effect sizes of all digital interventions combined compared to the control conditions ranged from g=0.3 to 1.41 for all outcomes together (median g=0.78). Of the 20 effect sizes we extracted (Table 2), 4 were small (g 0.21-50), 9 were moderate (g 0.51-0.80), and 7 were large (g>0.81). The effect sizes we found for social anxiety disorder in these meta-analyses ranged from g = 0.58 to 0.79 (median g = 0.76), for generalized anxiety disorder from g = 0.3 to 1.06 (median g = 0.67), and for panic disorder anxiety from g = 0.91 to 1.41 (median g = 1.15). Specific outcomes for mixed anxiety ranged from g = 0.3 to 0.32 (median g = 0.31). Specific outcomes for other anxiety symptoms, like worry or specific phobias, ranged from g = 0.75 to

1.07 (median $g = 0.91$). All effect sizes ofr primarily internet-based interventions were statistically significant, but the effect size for smartphone applications were not (Firth et al., 2017; Linardon et al., 2019; Weisel et al., 2019).

Heterogeneity was high for most effect sizes. Of the 20 extracted effect sizes that reported I^2, only 2 (10%) reported a value of I^2 below 50%, 10 between 50 and 75%, and 7 (35%) reported a value above 75%.

3.1.3. Characteristics of the 62 included primary studies

The total number of primary trials that were included in the 10 meta-analyses was 62, a total number of 188 effect sizes (Table 2). Of these trials, 26 were identified by only one meta-analysis (unique primary studies; 13.83%).

We were able to extract a quality score for 43 studies (69.35% of studies; Table 2). We approximated the quality ratings from primary study level summary ratings, i.e low, medium, high risk of bias for 13 studies, and could only obtain overall quality rating scores on a meta-analytic level for two meta-analyses containing 6 studies.

Only 10 primary studies had a score of 0.90 or higher (17.86% of the studies with a quality score), and 19 (33.93%) had a score above 0.7.

Most included primary studies investigated the efficacy of digital interventions delivered via the internet (87.23%). Most studies used guided interventions (72.87%) and were primarily targeted at generalized anxiety disorders (20.74%). Almost all study participants were recurited from the community (94.68 %). Wait list control groups were the most prevalent comparison group (96.28%). Only 26.06% of studies were rated with a low risk of bias. See Table 3 for a summary of all primary study characteristics.

Table 3. *Summary characteristics of included effect sizes from primary studies.*

Characteristics	***Umbrella Review (N = 188[1])***	***Multiverse Meta-Analysis (N = 238[2])***
Origin		
Database 2023	-	50 (21%)
Domhardt 2020	22 (12%)	22 (9.2%)
Eilert 2021	21 (11%)	21 (8.8%)
Firth 2017	3 (1.6%)	3 (1.3%)
Guo 2020	14 (7.4%)	14 (5.9%)
Linardon 2019	7 (3.7%)	7 (2.9%)
Mor 2021	6 (3.2%)	6 (2.5%)
Pauley 2021	53 (28%)	53 (22%)
Romijn 2019	41 (22%)	41 (17%)
Stech 2019	16 (8.5%)	16 (6.7%)
Weisel 2019	5 (2.7%)	5 (2.1%)
Tech		
Internet	164 (87%)	199 (84%)
Smartphone	14 (7.4%)	29 (12%)
Internet And Smartphone	10 (5.3%)	10 (4.2%)
Guidance		
Both	5 (2.7%)	5 (2.1%)
Guided	137 (73%)	163 (68%)
Unguided	46 (24%)	70 (29%)
Diagnosis		
GAD	39 (21%)	58 (24%)
Mixed Anxiety	39 (21%)	40 (17%)
Other Anxiety	9 (4.8%)	13 (5.5%)
PDa	58 (31%)	68 (29%)
SAD	43 (23%)	59 (25%)
Recruitment		
Clinical	10 (5.3%)	12 (5.0%)
Community	178 (95%)	226 (95%)
Control		
Other Control	7 (3.7%)	16 (6.7%)
WLC	181 (96%)	222 (93%)
Risk of Bias		
High	34 (18%)	43 (18%)
Some Concern	105 (56%)	63 (26%)
Low	49 (26%)	63 (26%)

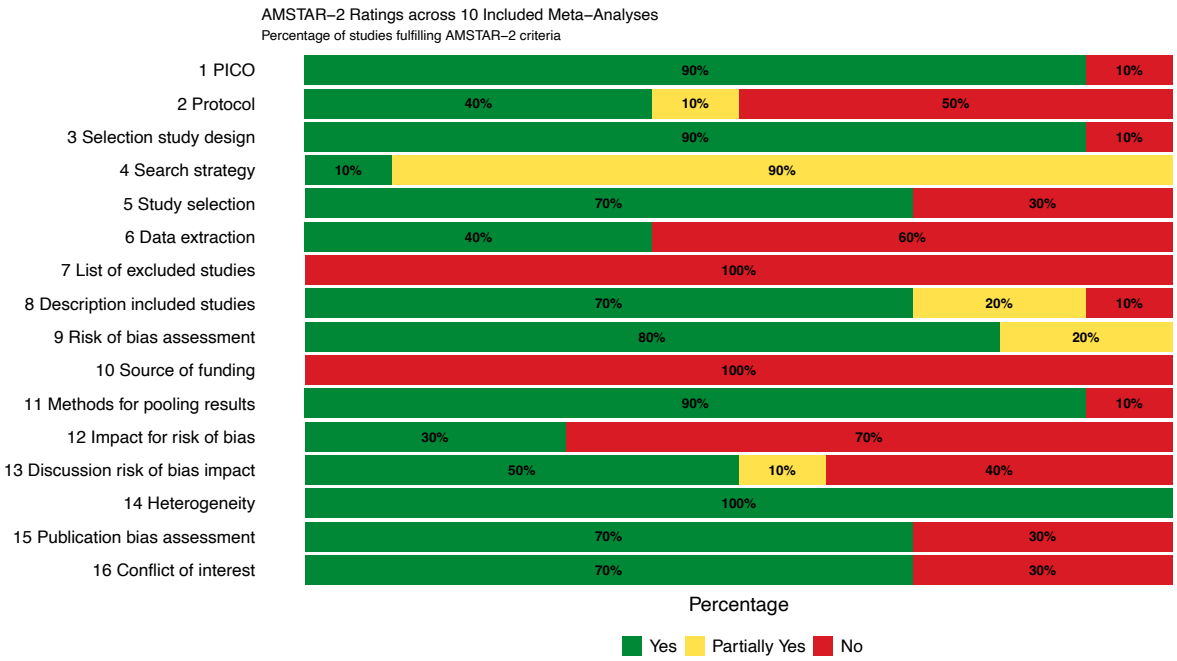
[1]Effect sizes from umbrella review including studies from 10 meta-analyses (%); [2]Effect sizes from multiverse meta-analyses that additionally include primary studies from targeted search.

3.2.1. AMSTAR-2 Ratings

Overall, 5 meta-analyses showed critically low AMSTAR-2 ratings, while 3 were rated as low, and 2 were rated as moderate. The AMSTAR-2 ratings for each of the 10 included meta-analyses are given in Table 2. No meta-analysis was rated positive for all 16 items, 2 meta-analyses were rated positive for 12 to 15 items, and 7 were rated positive for 8 to 11 items, and 1 meta-analysis achieved fewer than 8 points.

The aggregated ratings across all 10 meta-analyses are reported in Figure 2. Most meta-analyses (90%) reported the PICO correctly (item 1), and all meta-analyses included an explanation and discussion of heterogeneity (item 14). Three more items were rated positive by more than 80% of the meta-analyses (items 3, 9 and 11), while 3 items were rated positive in less than 10% of the meta-analyses (items 4, 7, and 10).

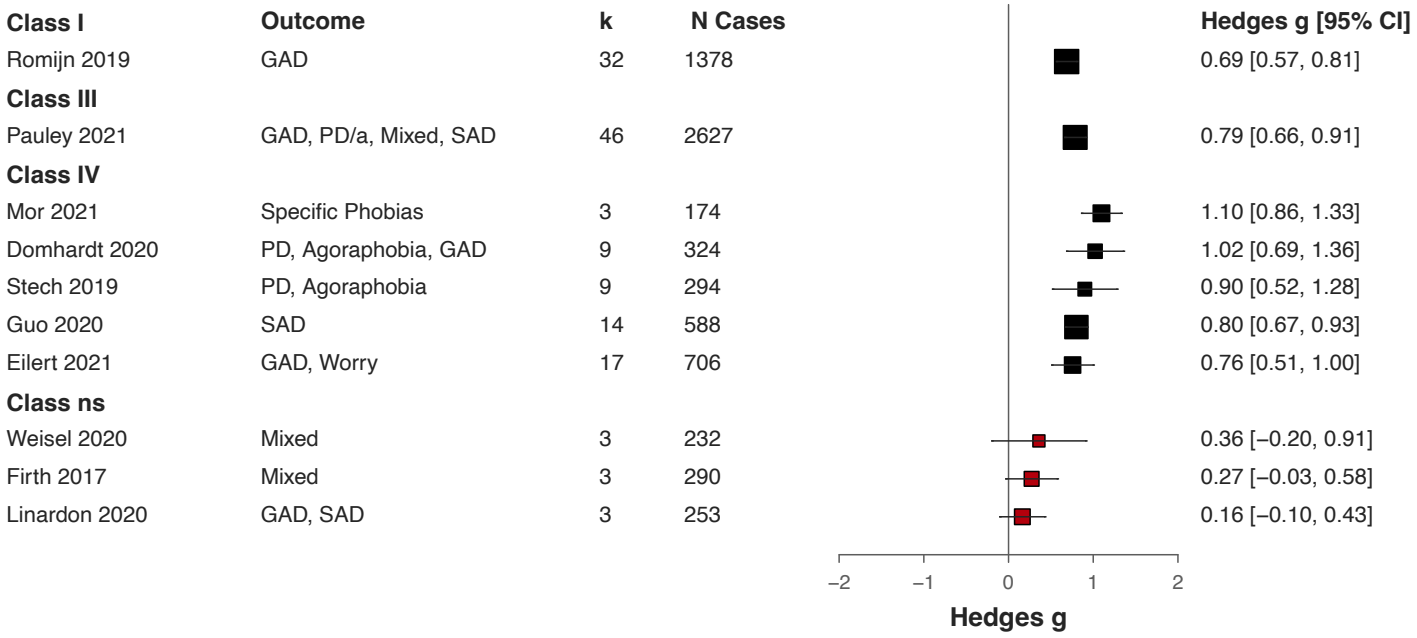
Figure 2. *AMSTAR-2 Ratings of all 11 Meta-Analyses.*



3.1.4. Strength of Evidence

Only one meta-analysis reached Class I (convincing evidence), and only one meta-analyses reached Class III (suggestive evidence). All other meta-analyses had a weak strength of evidence for the efficacy of digital interventions for anxiety disorders. Effect sizes larger than $g = 1$ were only seen in meta-analyses with weak evidence, while the two meta-analyses with higher evidence classes ranged from $g = 0.69$ to 0.79. See Figure 3 for a forest plot of all 10 meta-analyses.

Figure 3. *Strength of evidence of meta-analyses on digital interventions for anxiety disorders.*



Note. Class I (convincing evidence): number of cases > 1000, *p*-value of the meta-analysis < 10e-6, I^2 < 50%, 95% prediction interval excluding the null, *p*-value of the Egger's test > .05 and *p*-value of the Ioannidis' test > .05; Class II (highly suggestive evidence): number of cases > 1000, *p*-value of the meta-analysis < 10e-6, largest study with a statistically significant effect and class I criteria not met; Class III (suggestive evidence): number of cases > 1000, *p*-value of the meta-analysis < 10e-3 and class I–II criteria not met; Class IV (weak evidence): *p*-value of the meta-analysis < 0.05 and class I–III criteria not met; Class ns (not statistically significant): *p*-value of the meta-analysis >= 0.05

When calculating evidence classes for the different anxiety disorders, the strongest meta-analytic evidence was found for GAD (suggestive evidence, $g = 0.61$, 95% CI [0.34, 0.87]). For panic disorder and other anxiety types, the evidence was found to be weak (Figure A.1). When calculating evidence classes separately by type of control condition, the strongest meta-analytic

evidence was found for the efficacy of digital interventions compared with waitlist control groups (Figure A.2).

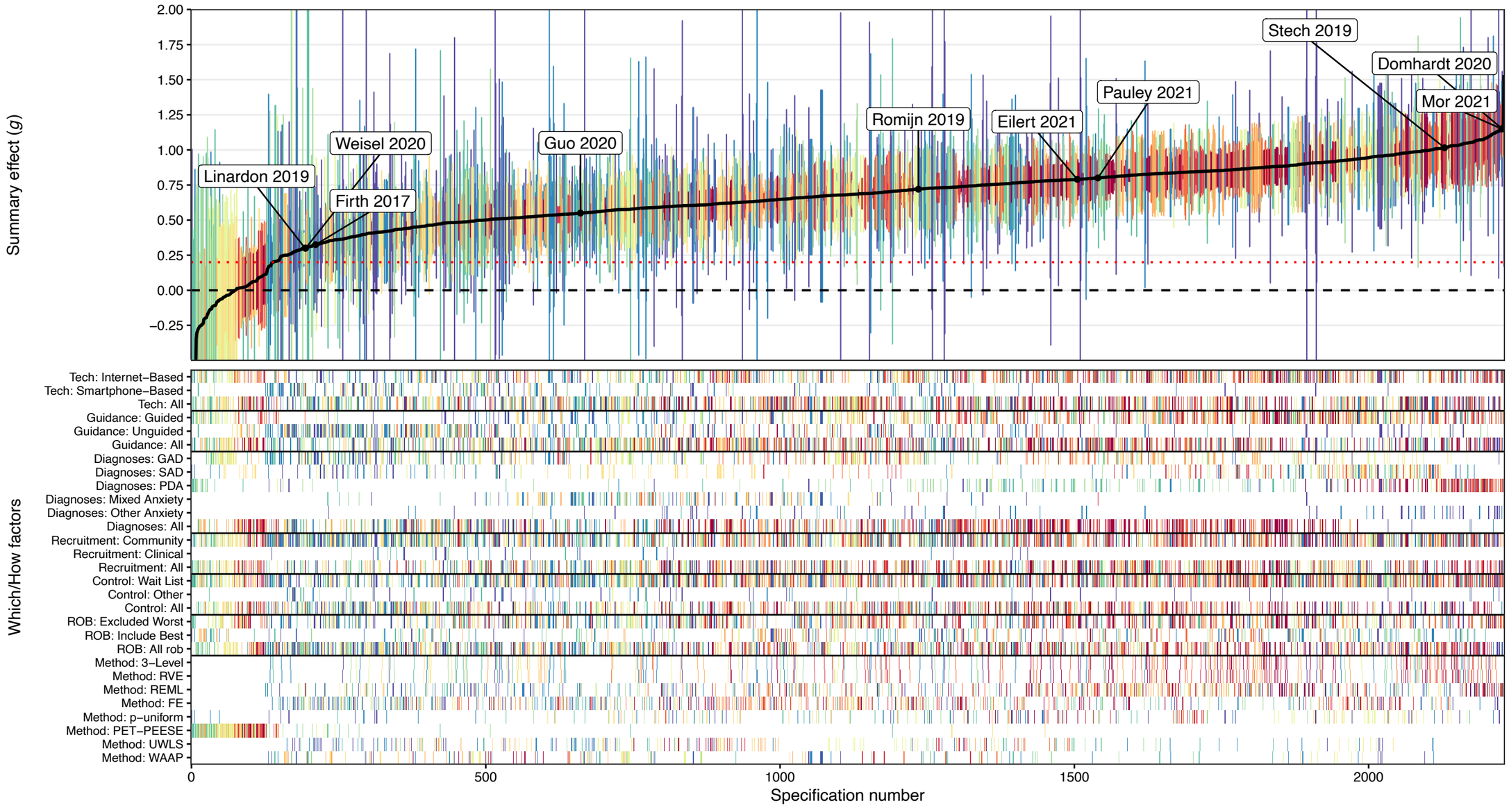
3.2.Multiverse Meta-Analysis

3.3.1. Descriptive Specification Curve

The descriptive specification curve showed that the summary effect sizes of meta-analyses can vary from null effects to very large effect sizes. See Figure 6 for a detailed visualization of all 2220 meta-analyses based on all combinations of *Which* and *How* factors in 81 randomized controlled trials. The estimated mean Hedges' *g* of the various subsets ranged from -0.75 to 1.53, with an interquartile range of 0.52 to 0.84. In total, 96.67 of the estimated means were greater than 0, and 86.76% of these had 95% *CIs* that did not include 0 (i.e., estimated means greater than 0 which would have returned a two-tailed *p*-value of less than .05).

In total, 93.78% reached a small effect size of Hedges' $g > 0.2$, and 77.61% of the summary effect sizes had 95% *CIs* larger than a small effect size.

Figure 4. *Descriptive Specification Curve Plot of all possible meta-analyses on digital interventions for anxiety disorders*



Note. The top panel shows the meta-analytic summary effects (*g*) for each specification with their 95% confidence interval. The summary effects are sorted by their magnitude, ranging from lower to higher. Connecting the different summary effects results in the solid line, which is the specification curve. A horizontal dashed line of no effect is shown at $g = 0$ and a red dotted line is shown to indicate a small, yet clinically relevant effect size at $g = 0.3$. Published meta-analyses are shown on their respective position on the specification curve. The vertical columns in the bottom panel represent factor combinations of all *Which* factors. These include different technologies, types of guidance, diagnoses, recruitment strategies, control conditions, and strategies to deal with the risk of bias. The *How* factors include several meta-analytical estimators: 3-level meta-analytical models, RVE = robust variance estimation, REML = restricted maximum likelihood estimation, FE = fixed effect model, *p*-uniform*, PET-PEESE, UWLS, and WAAP). Each vertical column is color-coded, signifying the number of samples included in a specification (hot spectral colors for more included samples vs. cool spectral colors for less included samples). The overall pattern of the specification curve indicates that specifications with many samples and/or narrow confidence intervals are closer to the interquartile range of estimated means as opposed to specifications with only a few samples and/or wider confidence intervals.

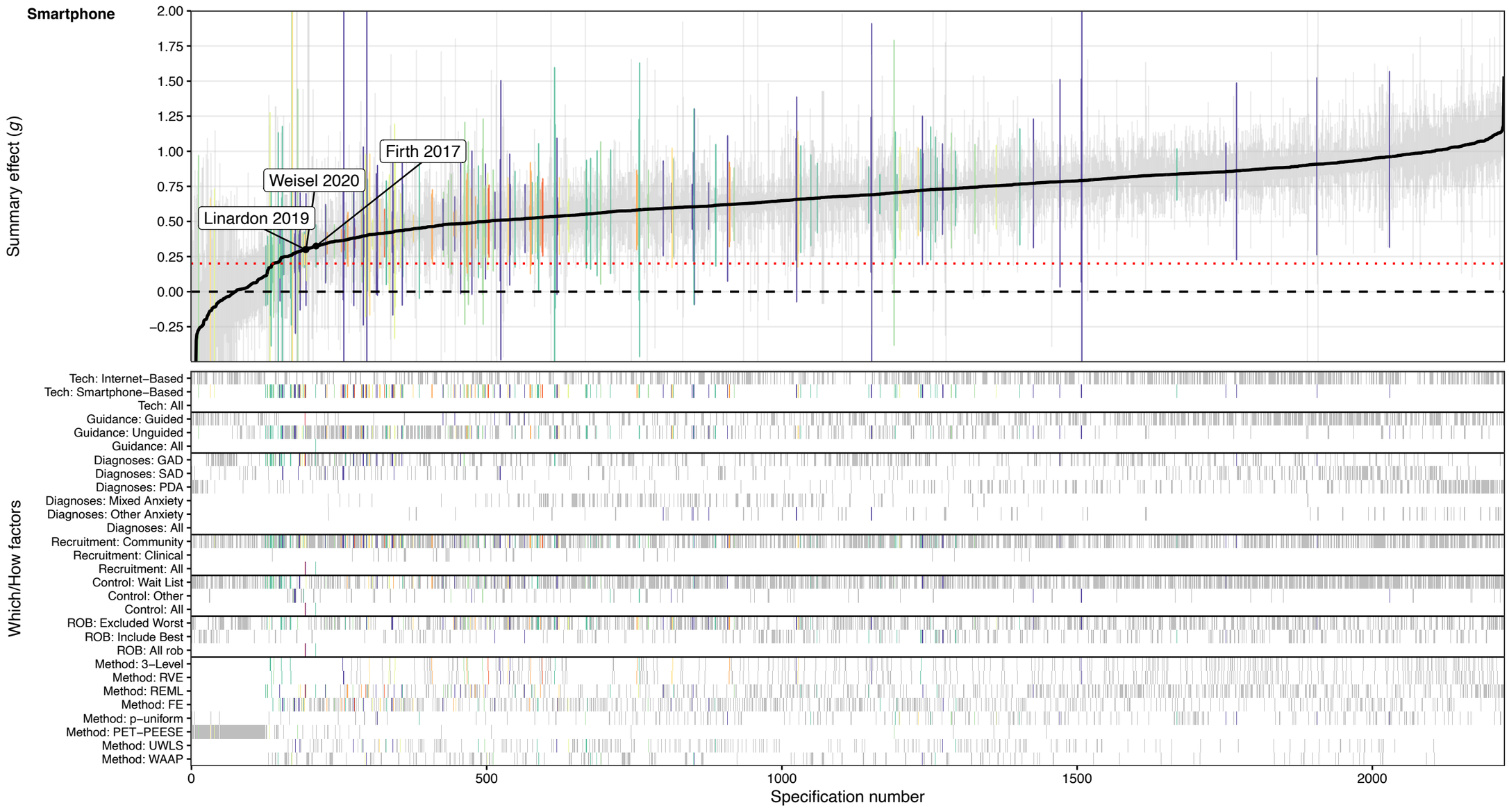
3.2.2. *Which* factors systematically associated with the magnitude

Some *Which* factors seemed to be systematically associated with the magnitude of the summary effect while other *Which* factors did not.

3.2.2.1. Technology of intervention

We found different summary effect sizes depending on the type of utilized technology for the intervention. On average, meta-analyses investigating internet interventions, mean $g = 0.69$, 95% *CI* [0.42, 0.95] with $k = 863$ included meta-analyses, produced larger effect size estimates than meta-analyses investigating smartphone-based interventions, mean $g = 0.48$, 95% *CI* [0.09, 0.87] with $k = 185$. See Figure A.2. for differences in summary effect sizes grouped by the technology of the intervention and Figure A.3. for a descriptive specification curve plot focusing on smartphone-based interventions.

Figure 5. *Descriptive specification curve highlighting smartphone interventions*



Note. The top panel shows the meta-analytic summary effects (*g*) for each specification with their 95% confidence interval. The summary effects are sorted by their magnitude, ranging from lower to higher. Connecting the different summary effects results in the solid line, which is the specification curve. A horizontal dashed line of no effect is shown at $g = 0$ and a red dotted line is shown to indicate a small, yet clinically relevant effect size at $g = 0.3$. Published meta-analyses are shown on their respective position on the specification curve. The vertical columns in the bottom panel represent factor combinations of all *Which* factors. These include different technologies, types of guidance, diagnoses, recruitment strategies, control conditions, and strategies to deal with the risk of bias. The *How* factors include several meta-analytical estimators: 3-

level meta-analytical models, RVE = robust variance estimation, REML = restricted maximum likelihood estimation, FE = fixed effect model, *p*-uniform*, PET-PEESE, UWLS, and WAAP). Each vertical column is color-coded, signifying the number of samples included in a specification (hot spectral colors for more included samples vs. cool spectral colors for less included samples). The overall pattern of the specification curve indicates that meta-analyses including only smartphone-based interventions are smaller compared to specifications including other interventions.

3.2.2.2.Guidance

Summary effect sizes differed depending on the type of guidance utilized in the intervention. Meta-analyses that only included guided interventions, mean $g = 0.74$, 95% *CI* [0.49, 0.98] with $k = 705$, produced larger effect size estimates than meta-analyses including only unguided interventions, mean $g = 0.51$, 95% *CI* [0.15, 0.87] with $k = 388$. See Figure A.4 for differences in summary effect sizes grouped by the type of guidance and Figure A.5. for a descriptive specification curve plot focusing on unguided interventions.

3.2.2.3.Anxiety Disorders

The type of anxiety disorder that was investigated in the analyses was found to have a substantial impact on the summary effect sizes. Meta-analyses that included samples assessing panic disorder with or without agoraphobia, mean $g = 0.8$, 95% *CI* [0.48, 1.13] with $k = 267$ included samples, produced larger effect size estimates than meta-analyses assessing generalized anxiety disorder, mean $g = 0.58$, 95% *CI* [0.25, 0.91] with $k = 496$. See Figure A.6. for differences in summary effect sizes grouped by the type of anxiety disorder and Figure A.7. for a descriptive specification curve plot focusing on PD/A.

3.2.2.4.Recruitment

The recruitment strategy did barely produce differences in effect size estimates. However, only very few meta-analyses could exist based on purely clinical populations. Meta-analyses that included samples recruited from the community, mean $g = 0.64$, 95% *CI* [0.34,

0.94] with k = 1260, produced larger effect size estimates than meta-analyses based on clinical samples, mean g = 0.5, 95% *CI* [0.20, 0.81] with k = 42. See Figure A.8. for differences in summary effect sizes grouped by the type of recruitment and Figure A.9. for a descriptive specification curve plot focusing on recruitment from clinical populations.

3.2.2.5.Control Group

Summary effect sizes differed depending on the type of control group utilized in the meta-analyses. Meta-analyses that compared interventions with a wait list control group, mean g = 0.67, 95% *CI* [0.39, 0.95] with k = 1281 included samples, produced larger effect size estimates than meta-analyses that compared interventions with other control groups, mean g = 0.57, 95% *CI* [-0.01, 1.15] with k = 97 included samples. See Figure A.10. for differences in summary effect sizes grouped by the type of control group and Figure A.11. for a descriptive specification curve plot focusing on waitlist control groups.

3.2.2.6.Different Strategies in Dealing with the Risk of Bias of Primary Studies

The different choices in dealing with high risk of bias did not result in different effect size estimates. Meta-analyses that included only high-quality/low risk of bias studies, mean g = 0.68, 95% *CI* [0.36, 0.98] with k = 338 had very similar effect sizes as meta-analyses that included all studies, mean g = 0.63, 95% *CI* [0.33, 0.93] with k = 1057. See Figure A.12. for differences in summary effect sizes grouped by the strategy of dealing with risk of bias and Figure A.13. for a descriptive specification curve plot focusing on including only low risk of bias studies.

3.2.3. How Factors

Among the meta-analytical methods we used, most yielded very similar effect size estimates, except PET-PEESE, a method designed to correct for publication bias. Robust variance estimation, for example, produced larger effect size estimates (mean $g = 0.76$, 95% *CI* [0.30, 0.121], $k = 251$) compared to PET-PEESE (mean $g = 0.01$, 95% *CI* [-0.55, 0.58], $k = 149$).

Looking at the variation in confidence intervals across the different methods, a similar picture emerged. Those methods that correct for publication bias (PET-PEESE, *p*-uniform, WAAP, UWLS) resulted in a statistically larger effect size estimate than a small effect size of Hedges' $g = 0.2$ in 61% of meta-analyses. On the other hand, methods that did not correct for publication bias produced distinctly larger effect size estimate in 86% of meta-analyses, suggesting that non-correcting methods tend to produce larger effect sizes.

Similarly, methods that ignored effect size dependency produced effect size estimates that surpassed a small effect size of 0.2 (with a confidence interval larger than this effect size estimate) in 90% of meta-analyses. Conversely, methods that aggregated or modeled the effect size dependency exceeded this small effect size of $g = 0.2$ only in 72% and 77% of meta-analyses, respectively.

4. Discussion

4.1.Key Findings and Contributions of the Study

We critically evaluated and reviewed a substantial amount of research on digital interventions for anxiety disorders from 10 meta-analyses, 81 primary studies and 248 effect sizes. The included meta-analyses reported small but statically not significant effect size estimates for smartphone-based interventions and moderate-to-large effects for guided, self-guided, and internet-based interventions. In a sensitivity check, we investigated the efficacy of digital interventions for anxiety by simultaneously analyzing all possible meta-analyses (k =2220) that could be conducted based on all reasonable combinations of inclusion criteria and meta-analytical methods. We found that most meta-analyses produced small to large effect sizes, suggesting the overall robustness of findings. With the current body of evidence, most possible meta-analyses would find effect sizes statistically different from 0. Although the exact effect sizes differed substantially between both the published and all estimated meta-analyses, it seems safe to conclude that some of the available interventions can be effective in reducing the symptoms of anxiety.

However, the quality of most meta-analyses was far from optimal. Only 2 of the 11 meta-analyses rated positive for 12 or more of the 16 items of the AMSTAR-2, and only one meta-analysis attained the highest level of strength of the evidence attained—all other meta-analyses provided only weak or not significant evidence. These low-quality scores make the results of the meta-analyses uncertain. The quality of the primary studies was sub optimal as well, as only 26% obtained a low risk of bias rating. Based on the data provided in the meta-analyses, we could calculate a standardized quality score for two thirds of the primary studies (69%) and had to approximate the rest. Only 18% of primary studies had a risk of bias score of 0.9 or higher on a

scale from 0 to 1, and only 29% scored above 0.75. Only half of the primary studies (50%) scored above 0.50, meaning that most primary studies are potentially biased and report inflated effect sizes.

The results from our multiverse meta-analyses suggest that the evidence is robust in regard to many methodological choices. The effects are very similar if we choose to only include the very best studies, if we exclude the worst studies, or if we decide to include all studies in the meta-analyses. Regarding the correction for publication bias, we found very similar effect sizes from a range of correction methods. The estimated efficacy of digital interventions was not strongly associated with the type of meta-analytical estimation method (except PET-PEESE correcting for publication bias but not with other methods correcting for publication bias), or the exclusion of high risk of bias studies. The magnitude of the meta-analyses was especially large in meta-analyses comparing a treatment with a waitlist control groups, when the intervention was targeted at individuals with panic disorder, when the intervention was delivered via the internet—contrasted to smartphone-based interventions— and when the intervention was guided compared to self-guided. The identified association that waitlist control groups produced larger effect sizes might be indicative of their relatively inferior effectiveness as an intervention. It is noteworthy that 93% of all included trials in the multiverse meta-analysis utilized a waitlist control group. This high prevalence is an essential aspect to consider because the over-reliance on waitlist controls could have led to an overestimation of treatment efficacy.

Overall, the observed discrepancies between meta-analyses contained both in the umbrella review and the multiverse meta-analysis can be primarily attributed to factors including the type of the intervention, the nature of the comparison group, and the assessed outcome. However, existing information is not sufficient to comprehensively clarify these discrepancies

for different populations (such as from clinical settings). Moreover, our analysis indicates that the different methodological choices made in the different meta-analyses did not seem to exert a significant influence.

The outcomes of both our umbrella review and multiverse meta-analysis cohesively point towards the same direction. The umbrella review, based on 10 meta-analyses, 62 unique primary studies, and 188 effect sizes, provided evidence for the effectiveness of internet-based interventions, and interventions compared to wait-list control groups. However, no significant evidence was found for smartphone-based interventions or when compared to active control groups. Complementing these findings, the multiverse analysis, involving 81 primary studies, 238 effect sizes, and 2220 meta-analyses, echoed similar trends. It revealed that smartphone interventions yielded smaller effect sizes compared to internet-based interventions, guided interventions were superior to unguided ones, and comparisons with WLC produced larger effect sizes. These consistent findings across both comprehensive reviews solidify our understanding of digital interventions' effectiveness in various settings and formats.

Considering our converging findings from both the umbrella review and the multiverse meta-analysis that smartphone-based studies are not only limited in number but also show less effectiveness, we identified a significant gap in our current understanding and implementation of these digital interventions. This gap could reflect a need for improved app design and functionality, better study design, or more robust theoretical underpinnings. However, our findings could also suggest a more fundamental issue - that smartphone-based interventions simply may not be as effective in treating anxiety disorders as internet-based interventions. Despite the ubiquity and convenience of smartphone apps, it might be possible they lack the necessary depth or the personalized approach needed to address such complex mental health

conditions effectively. This finding challenges the field to critically assess the utility and limits of smartphone-based digital therapeutics and underscores the importance of unified research efforts to identify digital interventions that can be effectively implemented.

4.2.Comparison of Study Results to Literature Findings

These results are predominantly in line with findings reported in two previous meta-reviews on smartphone-based interventions for mental health (Lecomte et al., 2020; Goldberg et al., 2022) and one meta-review on internet interventions for anxiety and mood disorders (Andersson et al., 2019). The main differences might have arisen due to different inclusion criteria for the meta-analyses and overall, more conservative quality ratings in our umbrella review.

While our main results overlap with Lecomte et al. (2020), who reported small-to-medium effect sizes for smartphone interventions focusing on anxiety symptoms, and smaller effect sizes for stand-alone apps, we rated the quality of evidence contained in those meta-analyses as substantially lower using different criteria (Ioannidis criteria instead of GRADE criteria). We did not include two out of their four meta-analyses (Stratton et al., 2017; Versluis et al., 2016) because the majority of included primary studies (>75%) in those meta-analyses were not focused on individuals with elevated anxiety symptoms but rather assessed the efficacy of treatments in healthy populations or participants with elevated symptoms in other mental health problems like depression or stress. Further research should focus on whether including anxiety symptoms as a primary or secondary outcome would yield different results.

We found similar magnitudes of effect sizes to Goldberg et al. (2022), but with different assessments of the strength of evidence. For example, they reported highly suggestive evidence of small magnitude effects on anxiety in the general population (d = 0.32, Linardon et al., 2019) and among those with elevated symptoms (d = 0.45, Firth et al., 2017) when compared to inactive controls. Their criteria for highly suggestive evidence was a sample size over 1000 and a *p-value* < 10−6, while our criteria were much stricter, as the number of cases (in the intervention arm) had to be greater than 1000, p-value of the meta-analysis < 10e-6, I^2 < 50%, 95% prediction interval excluding the null, p-value of the Egger test > .05 and p-value of the Ioannidis test > .05. Similar to our assessments, they reported weak evidence of small magnitude effects of apps compared to inactive controls among those with elevated symptoms (d = 0.49, Weisel et al., 2019), and weak evidence of very small effects compared to active controls in the general population and among those with elevated symptoms (ds = 0.18 and 0.19, respectively Firth et al., 2017). Beyond the scope of our review, they reported weak or non-significant evidence of similar magnitude for ecological momentary interventions and apps compared to non-specific controls (ds = 0.30 to 0.43, Loo Gee et al., 2016) and suggestive evidence of small magnitude effects for meditation apps (d = 0.31, Gál et al., 2021).

A narrative umbrella review by Andersson et al. (2019) concluded that ICBT can be effective in treating adults with anxiety. Based on several identified meta-analyses, they reported support for the efficacy in treating panic disorder (Adelman et al., 2014; Andrews et al., 2018; Apolinário-Hagen, 2019; Olthuis et al., 2016), social anxiety disorder (Andrews et al., 2018; Kampmann et al., 2016), and generalized anxiety disorder (Andrews et al., 2018; Richards et al., 2015), with moderate to large average effect sizes overall. We did not include several of those meta-analyses because they did not match our inclusion criteria, as they were either published

before 2017 (Adelman et al., 2014; Kampmann et al., 2016; Olthuis et al., 2016; Richards et al., 2015) or were focused on comorbid anxiety disorders (Andrews et al., 2018; Păsărelu et al., 2017).

4.3.Strengths and Limitations

The main strength of this study is providing the birds-eye-perspective and quality assessment of the whole field of digital psychological interventions for anxiety disorders, both on a meta-analytic and primary study level. Another strength of our study is that we were able to identify gaps in the existing research on digital interventions for anxiety. While there are many primary studies available on panic disorder and agoraphobia, generalized anxiety disorder, and social anxiety disorder, there is insufficient evidence on the effectiveness of digital psychological interventions for specific phobias and other types of anxiety disorders. Additionally, there is a gap in research on individuals recruited from clinical populations. This may be due to the fact that such individuals are already receiving face-to-face interventions or have a higher disease burden, which may make them less willing or able to use digital interventions as an alternative.

Another significant strength of our study that warrants mention is its pioneering use of the multiverse meta-analysis method in the realm of digital interventions. Notably, through this rigorous method, we identified areas both of robustness and of variation due to researchers' methodological choices. Moreover, our study marks a unique milestone as the first to combine umbrella reviews with multiverse meta-analyses. This innovative merger enables a comprehensive synthesis of existing literature alongside a nuanced examination of outcome robustness across diverse meta-analytical models. Consequently, it offers a more detailed and reliable insight into the field of digital interventions, setting a precedent for future research methodologies and reinforcing the validity of our findings.

There are also limitations, however, that must be acknowledged. First, out of the availability of existing research, we focused on short-term effects of the interventions, and it is not clear what the long-term effects of most interventions are. Second, because of the suboptimal quality of most meta-analyses and primary studies, all results have to be interpreted with caution. Third, heterogeneity was considerable for many of the identified effect sizes, especially those in meta-analyses on all types of interventions. This strengthens the need to interpret the findings with caution. One more limitation is that we decided to include meta-analyses published after 2017 to contain the most updated evidence, which may have resulted in some missed information. Another limitation was the inconsistent overlap in quality ratings between meta-analyses and/or meta-reviews, which was compounded by the fact that different instruments were used to assess meta-analytic evidence which measured different aspects. Additionally, even when the same instrument was used, the same primary study was often rated differently, and we had to make judgment calls ourselves (choosing the more conservative rating). This issue may be exacerbated when research teams use their own unique instruments or modify established instruments. Finally, although SMDs of 0.2 are usually considered as small, 0.5 as medium and 0.8 as large, SMDs are still a statistical concept and their size does not necessarily reflect the clinical relevance of it (Cuijpers et al., 2014).

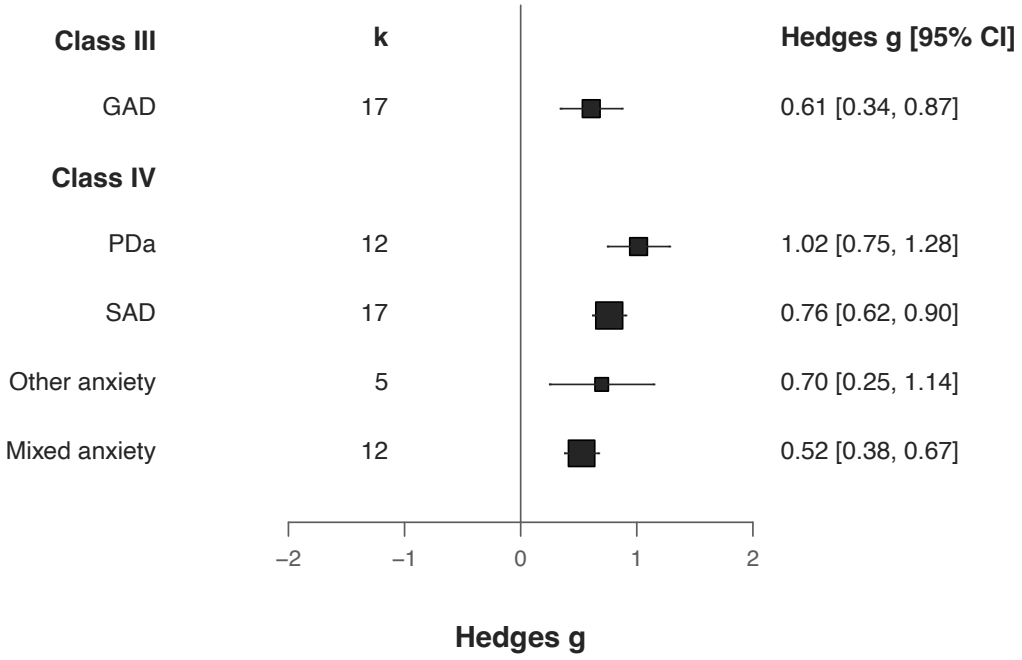
5. Conclusions

While our study suggests that digital interventions can be effective for treating anxiety disorders, caution should be exercised when interpreting these findings due to the low quality of studies included in our review. Additionally, both the presence and nontransparent reporting of conflicts of interest and financial interests in the field raise concerns about the objectivity of published studies. We found that waitlist control groups resulted in larger effect sizes, which

may reflect the poor quality of this control group as a comparison. It is also important to note that even-though our analyses to detect publication bias were inconclusive—publication bias is difficult to detect and still might have influenced the overall findings about the effectiveness of digital interventions. However, despite these limitations, the evidence overall suggests that digital interventions can be effective for treating anxiety disorders, with guided interventions showing greater efficacy compared to self-guided interventions and smartphone-based interventions being less effective than other types of digital interventions.

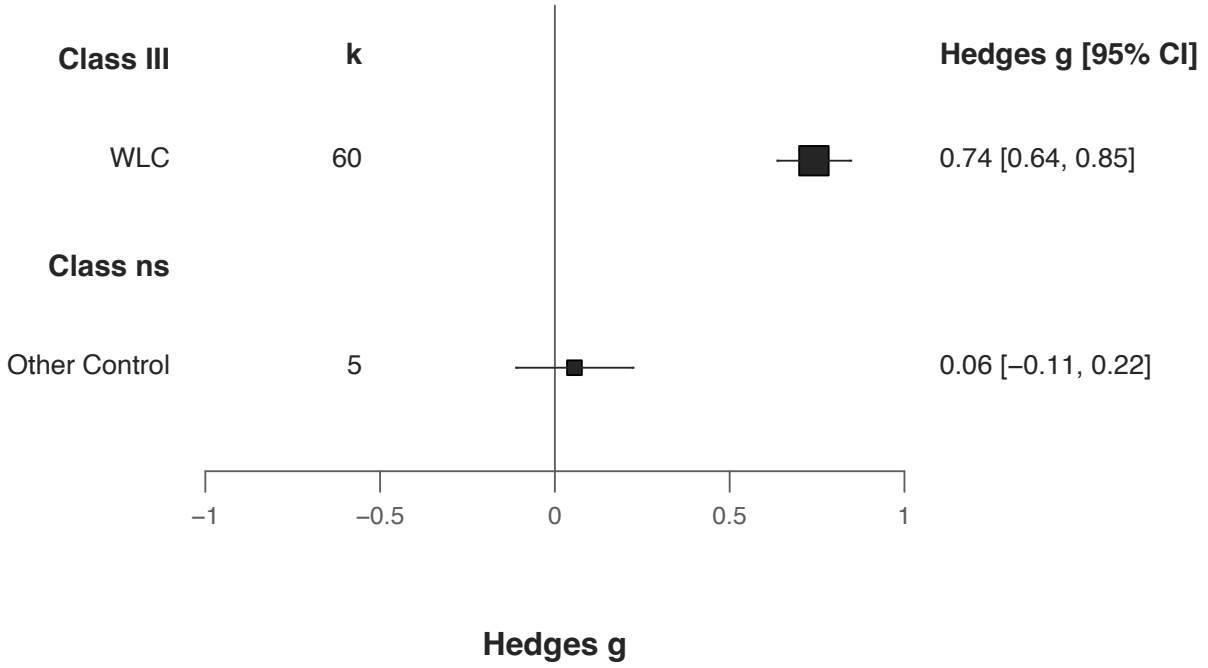
Appendix

Figure A.1. ***Strength of evidence of meta-analyses on digital interventions for different anxiety disorders.***



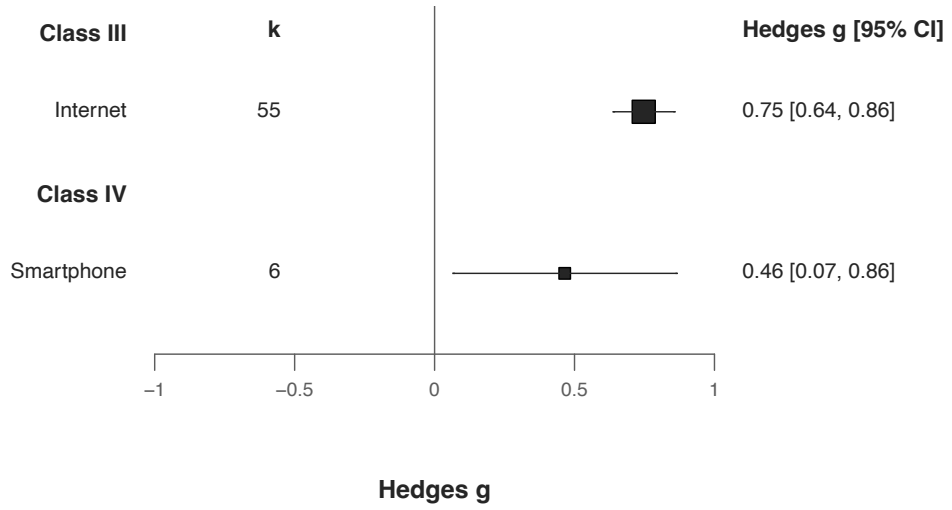
Note. Class I (convincing evidence): number of cases > 1000, *p*-value of the meta-analysis < 10e-6, I^2 < 50%, 95% prediction interval excluding the null, *p*-value of the Egger's test > .05 and *p*-value of the Ioannidis test > .05; Class II (highly suggestive evidence): number of cases > 1000, *p*-value of the meta-analysis < 10e-6, largest study with a statistically significant effect and class I criteria not met; Class III (suggestive evidence): number of cases > 1000, *p*-value of the meta-analysis < 10e-3 and class I–II criteria not met; Class IV (weak evidence): *p*-value of the meta-analysis < 0.05 and class I–III criteria not met; Class ns (not statistically significant): *p*-value of the meta-analysis >= 0.05

Figure A.2. ***Strength of evidence of meta-analyses on digital interventions for different control conditions.***



Note. Class I (convincing evidence): number of cases > 1000, *p*-value of the meta-analysis < 10e-6, I^2 < 50%, 95% prediction interval excluding the null, *p*-value of the Egger's test > .05 and *p*-value of the Ioannidis test > .05; Class II (highly suggestive evidence): number of cases > 1000, *p*-value of the meta-analysis < 10e-6, largest study with a statistically significant effect and class I criteria not met; Class III (suggestive evidence): number of cases > 1000, *p*-value of the meta-analysis < 10e-3 and class I–II criteria not met; Class IV (weak evidence): *p*-value of the meta-analysis < 0.05 and class I–III criteria not met; Class ns (not statistically significant): *p*-value of the meta-analysis >= 0.05

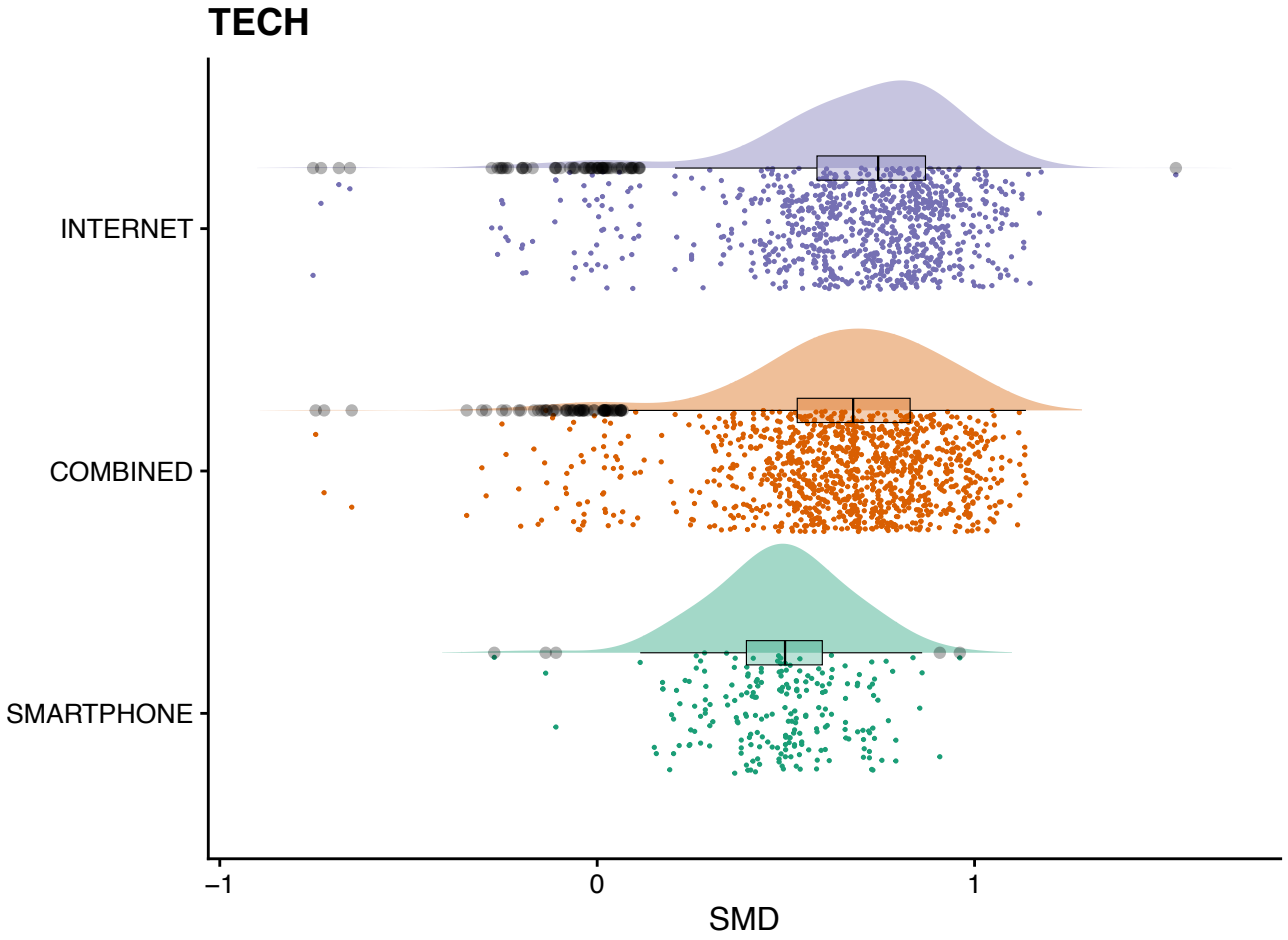
Figure A.3. ***Strength of evidence of meta-analyses on digital interventions for different technologies***



Note. Class I (convincing evidence): number of cases > 1000, *p*-value of the meta-analysis < 10e-6, I^2 < 50%, 95% prediction interval excluding the null, *p*-value of the Egger's test > .05 and *p*-value of the Ioannidis test > .05; Class II (highly suggestive evidence): number of cases > 1000, *p*-value of the meta-analysis < 10e-6, largest study with a statistically significant effect and class I criteria not met; Class III (suggestive evidence): number of cases > 1000, *p*-value of the meta-analysis < 10e-3 and class I–II criteria not met; Class IV (weak evidence): *p*-value of the meta-analysis < 0.05 and class I–III criteria not met; Class ns (not statistically significant): *p*-value of the meta-analysis >= 0.05

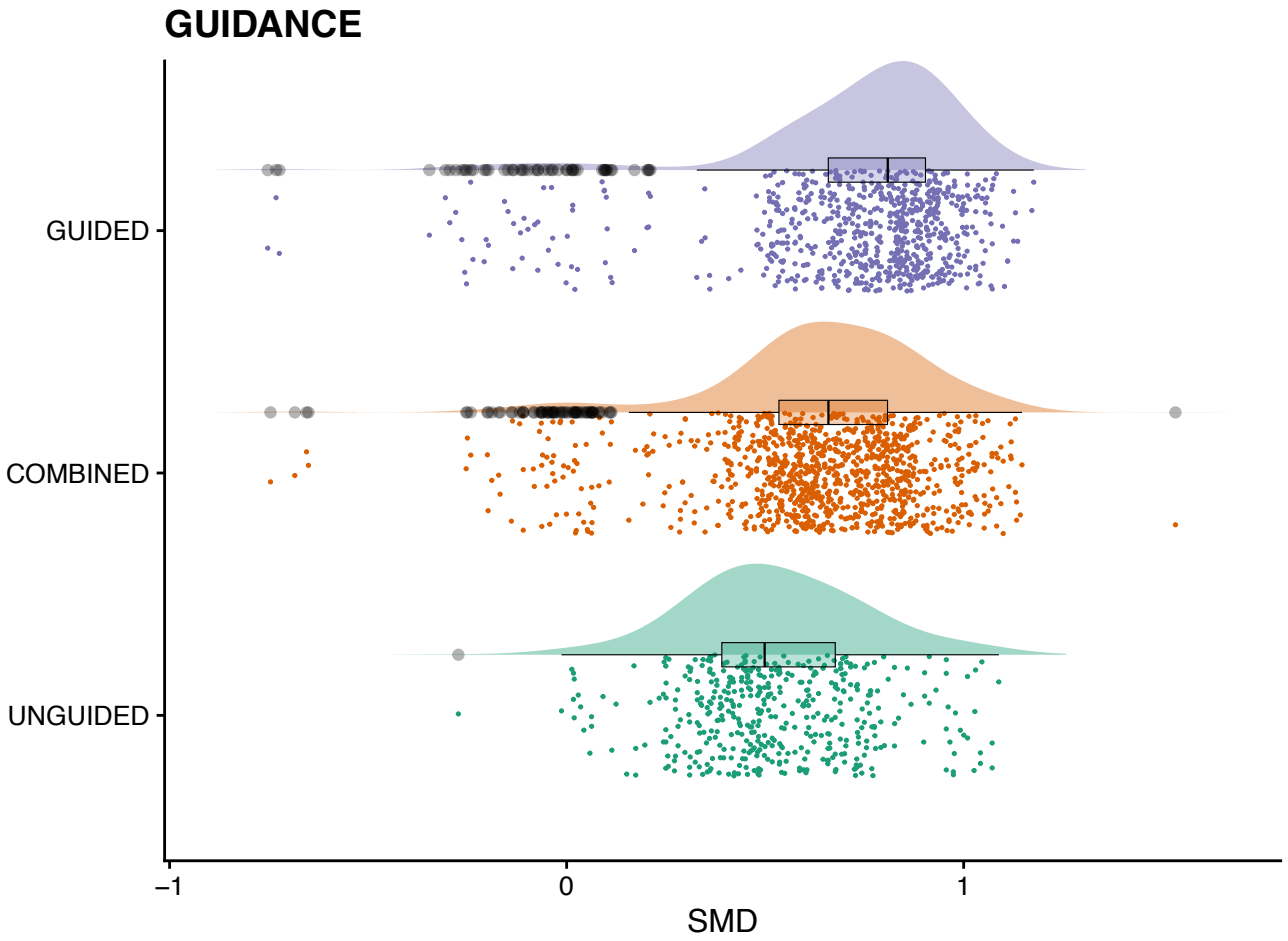
*Stolz 2018 and Ivanova 2016 excluded as they investigated mixed internet and smartphone based interventions.

Figure A.3. ***Raincloud plot of all meta-analyses on digital interventions for anxiety, grouped by technology of intervention.***



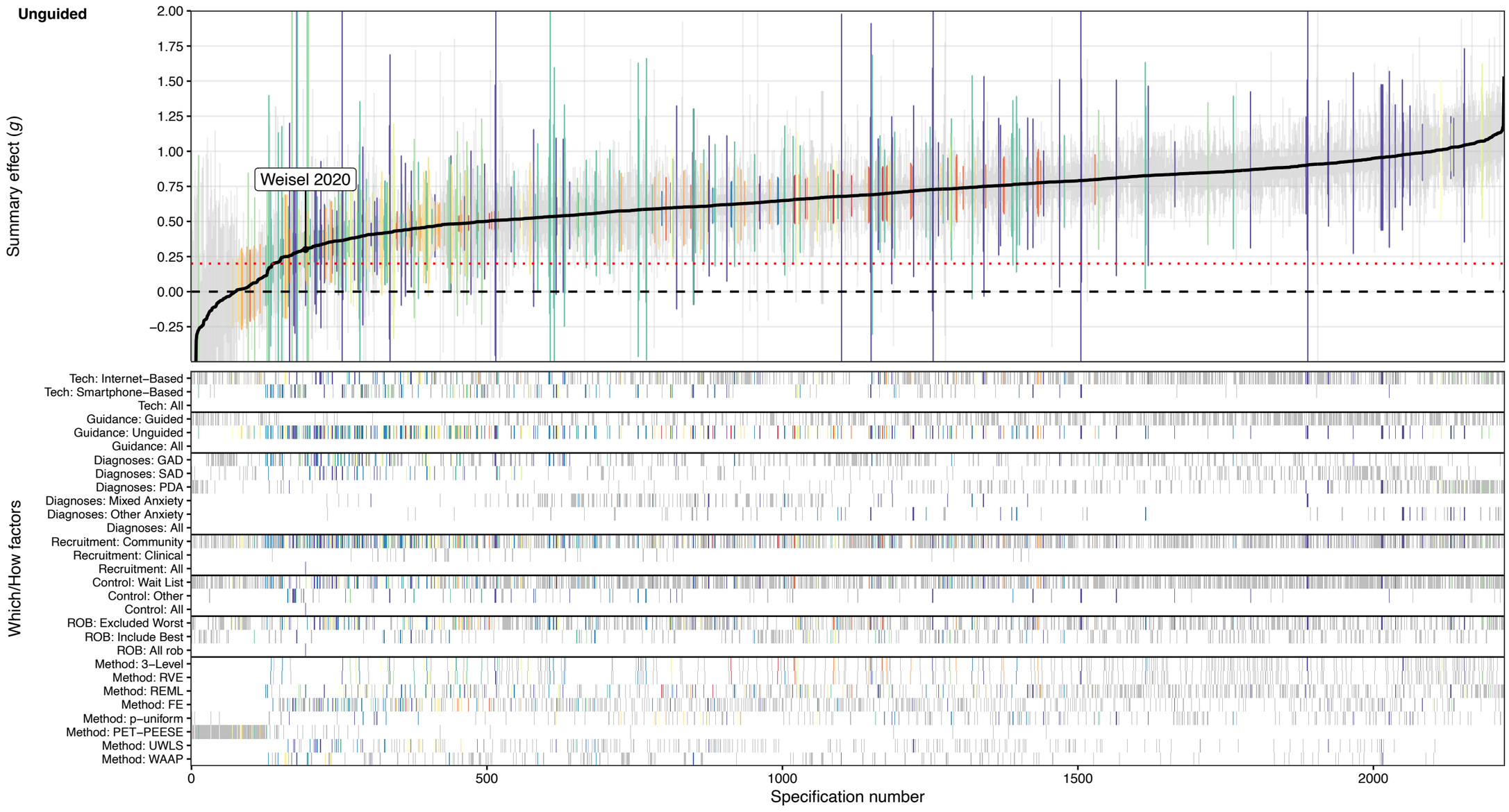
Note. Raincloud plots consist of three parts and depict the distribution of data (the cloud), a box-plot, and the raw data (the rain). They depict and visualize the distribution of summary effect sizes from all possible meta-analyses (produced by the multiverse meta-analysis) on internet-based, smartphone-based, and combined interventions.

Figure A.4. ***Raincloud plot of all meta-analyses on digital interventions for anxiety, grouped by type of guidance.***



Note. Raincloud plots consist of three parts and depict the distribution of data (the cloud), a box-plot, and the raw data (the rain). They depict and visualize the distribution of summary effect sizes from all possible meta-analyses (produced by the multiverse meta-analysis) on guided, self-guided, and combined interventions.

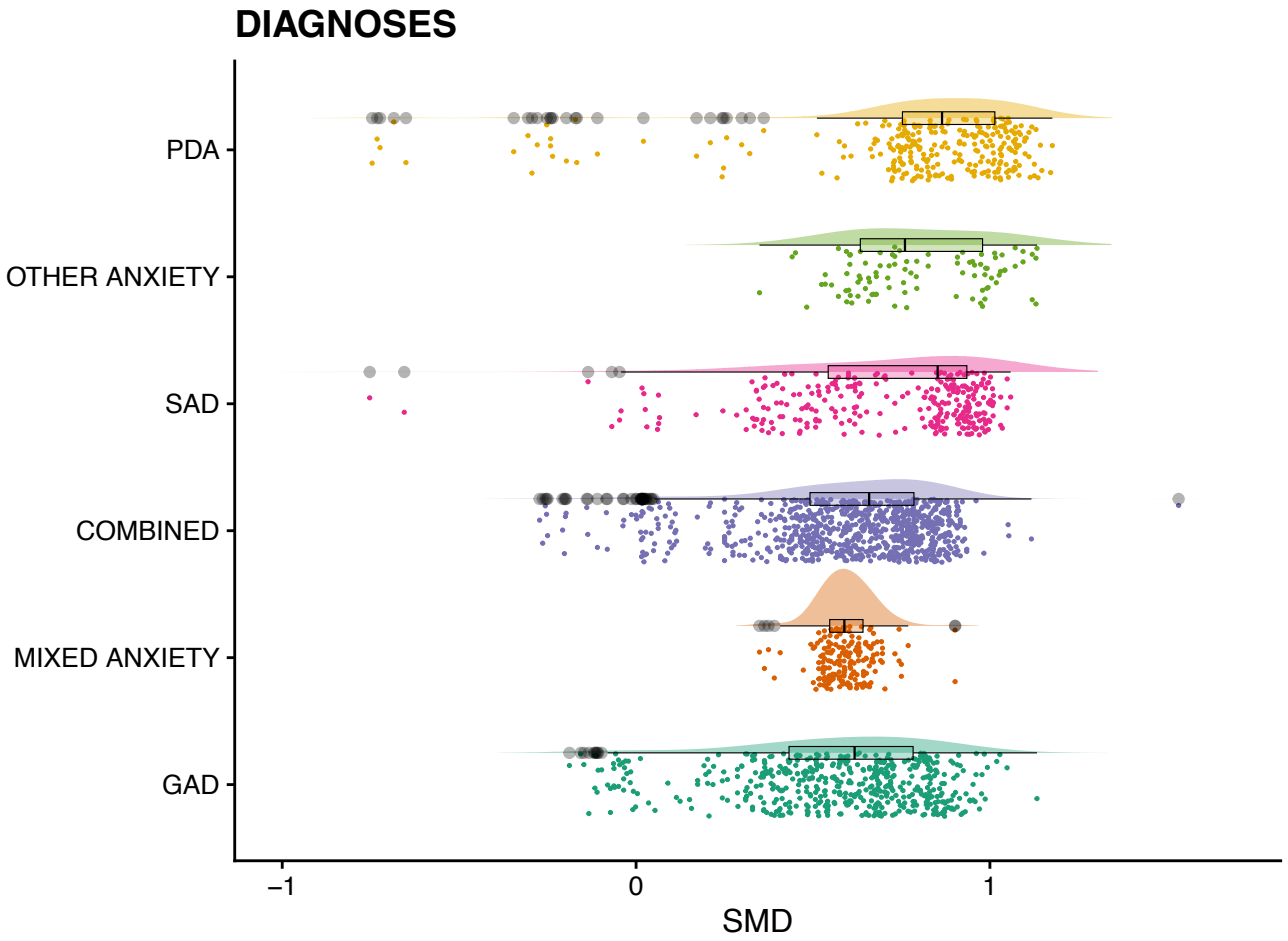
Figure A.5. ***Descriptive specification curve highlighting self-guided interventions.***



Note. The top panel shows the meta-analytic summary effects (*g*) for each specification with their 95% confidence interval. The summary effects are sorted by their magnitude, ranging from lower to higher. Connecting the different summary effects results in the solid line, which is the specification curve. A horizontal dashed line of no effect is shown at $g = 0$ and a red dotted line is shown to indicate a small, yet clinically relevant effect size at $g = 0.2$. Published meta-analyses are shown on their respective position on the specification curve. The vertical columns in the bottom panel represent factor combinations of all *Which* factors. These include different technologies, types of guidance, diagnoses, recruitment strategies, control conditions, and strategies to deal with risk of biases. The *How* factors include several meta-analytical estimators: 3-

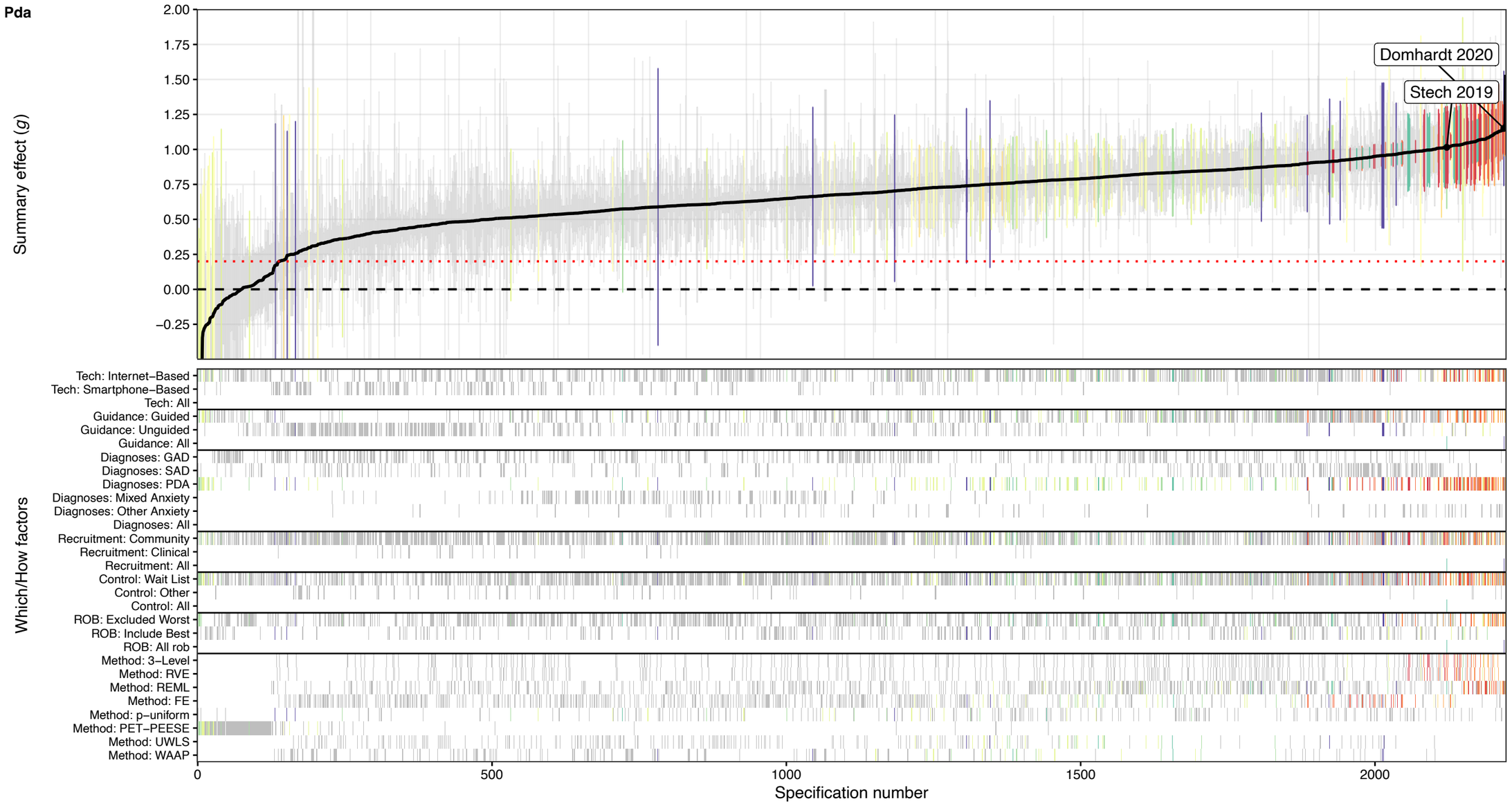
level meta-analytical models, RVE = robust variance estimation, REML = restricted maximum likelihood estimation, FE = fixed effect model, *p*-uniform*, PET-PEESE, UWLS, and WAAP). Each vertical column is color-coded, signifying the number of samples included in a specification (hot spectral colors for more included samples vs. cool spectral colors for less included samples). The overall pattern of the specification curve indicates that meta-analyses including only self-guided interventions have smaller effect sizes compared to specifications including guided interventions.

Figure A.6. ***Raincloud plot of all meta-analyses on digital interventions for anxiety, grouped by different anxiety disorders***



Note. Raincloud plots consist of three parts and depict the distribution of data (the cloud), a box-plot, and the raw data (the rain). They depict and visualize the distribution of summary effect sizes from all possible meta-analyses (produced by the multiverse meta-analysis) on all anxiety diagnoses (PDA = panic disorder with or without agoraphobia, SAD = Social anxiety disorder, GAD = generalized anxiety disorder, Mixed = a combination of anxiety disorders was assessed, Other Anxiety = other anxiety disorders, such as specific phobias or not specified).

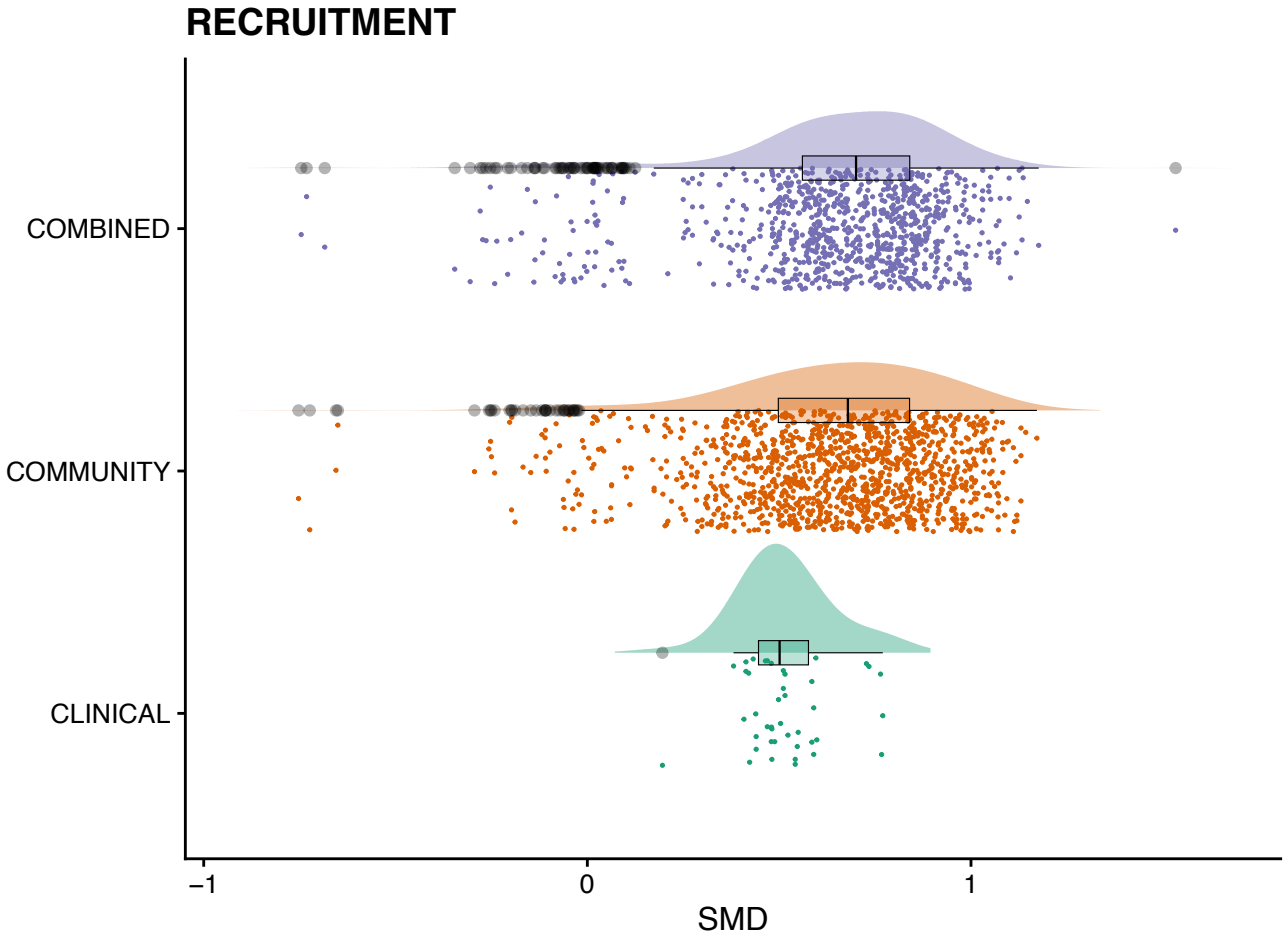
Figure A.7. ***Descriptive specification curve highlighting meta-analyses on panic disorder with or without agoraphobia***



Note. The top panel shows the meta-analytic summary effects (*g*) for each specification with their 95% confidence interval. The summary effects are sorted by their magnitude, ranging from lower to higher. Connecting the different summary effects results in the solid line, which is the specification curve. A horizontal dashed line of no effect is shown at $g = 0$ and a red dotted line is shown to indicate a small, yet clinically relevant effect size at $g = 0.2$. Published meta-analyses are shown on their respective position on the specification curve. The vertical columns in the bottom panel represent factor combinations of all *Which* factors. These include different technologies, types of guidance, diagnoses, recruitment strategies, control conditions, and strategies to deal with risk of biases. The *How* factors include several meta-analytical estimators: 3-

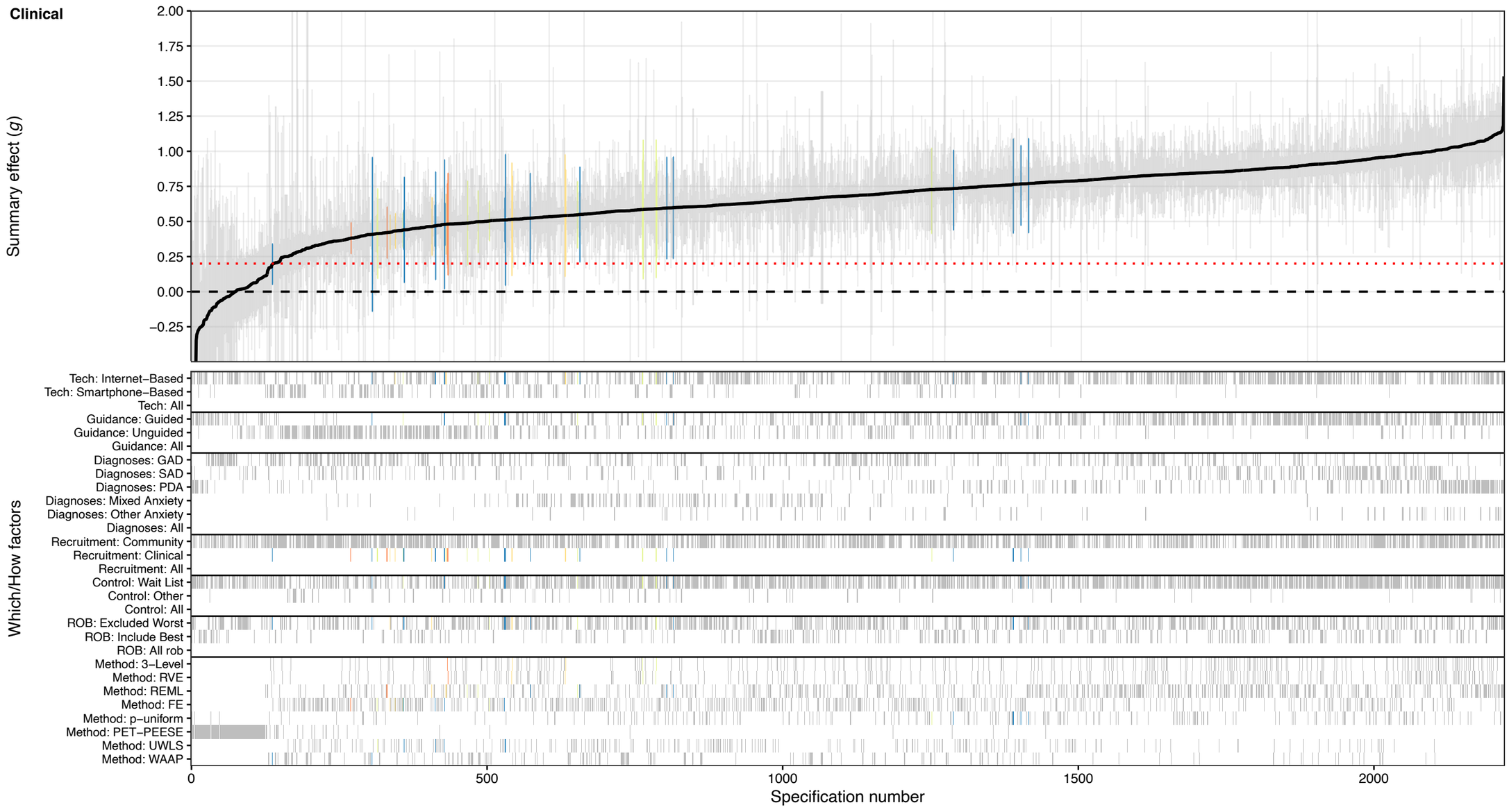
level meta-analytical models, RVE = robust variance estimation, REML = restricted maximum likelihood estimation, FE = fixed effect model, *p*-uniform*, PET-PEESE, UWLS, and WAAP). Each vertical column is color-coded, signifying the number of samples included in a specification (hot spectral colors for more included samples vs. cool spectral colors for less included samples). The overall pattern of the specification curve indicates that meta-analyses focusing on panic disorder have much larger effect sizes than those for other disorders.

Figure A.8. ***Raincloud plot of all meta-analyses on digital interventions for anxiety, grouped by different recruitment strategies***



Note. Raincloud plots consist of three parts and depict the distribution of data (the cloud), a box-plot, and the raw data (the rain). They depict and visualize the distribution of summary effect sizes from all possible meta-analyses (produced by the multiverse meta-analysis) having either a community sample, clinical sample, or combined sample.

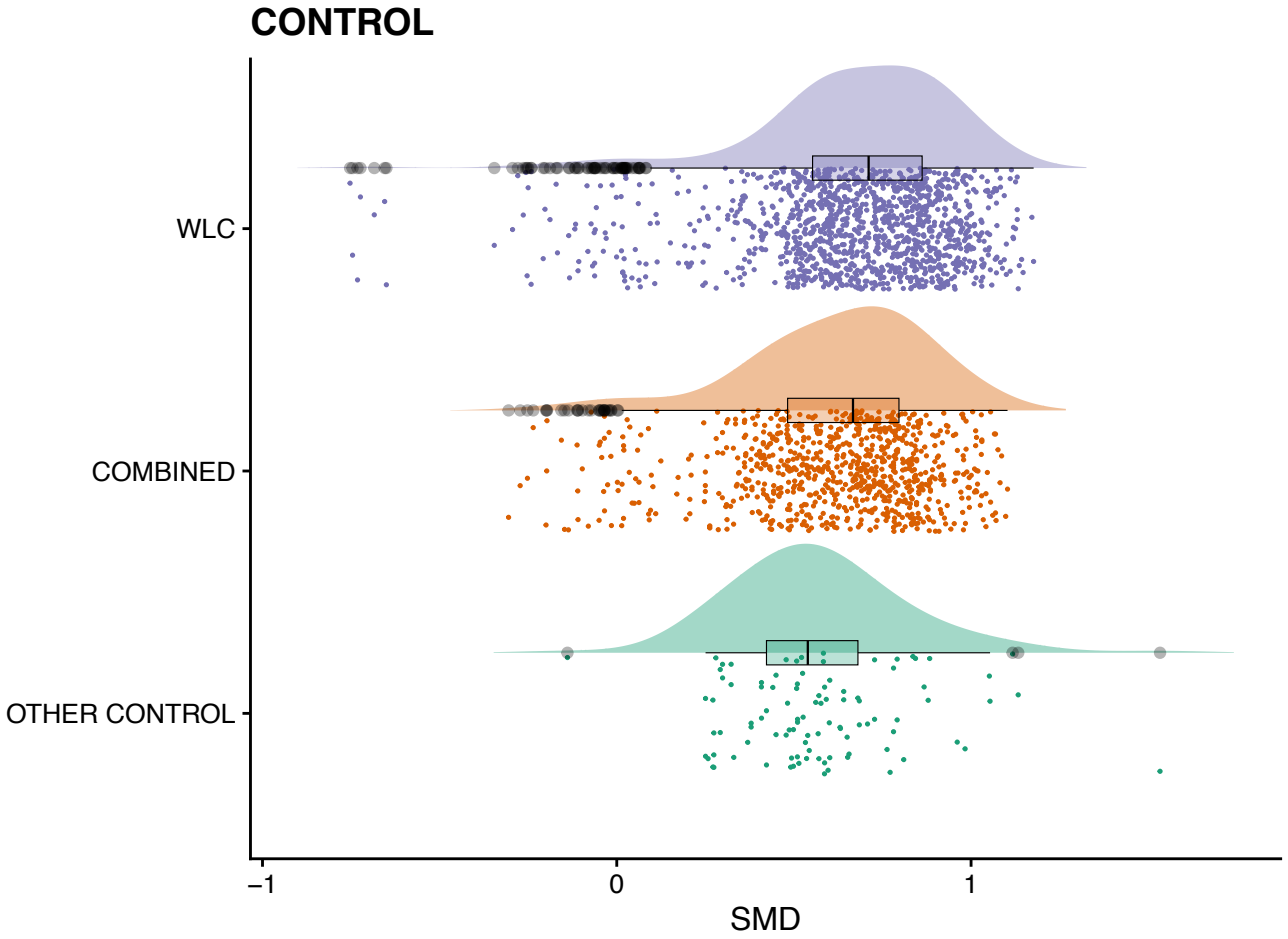
Figure A.9. ***Descriptive specification curve highlighting meta-analyses including only clinical samples***



Note. The top panel shows the meta-analytic summary effects (*g*) for each specification with their 95% confidence interval. The summary effects are sorted by their magnitude, ranging from lower to higher. Connecting the different summary effects results in the solid line, which is the specification curve. A horizontal dashed line of no effect is shown at $g = 0$ and a red dotted line is shown to indicate a small, yet clinically relevant effect size at $g = 0.3$. Published meta-analyses are shown on their respective position on the specification curve. The vertical columns in the bottom panel represent factor combinations of all *Which* factors. These include different technologies, types of guidance, diagnoses, recruitment strategies, control conditions, and strategies to deal with risk of biases. The *How* factors include several meta-analytical estimators: 3-

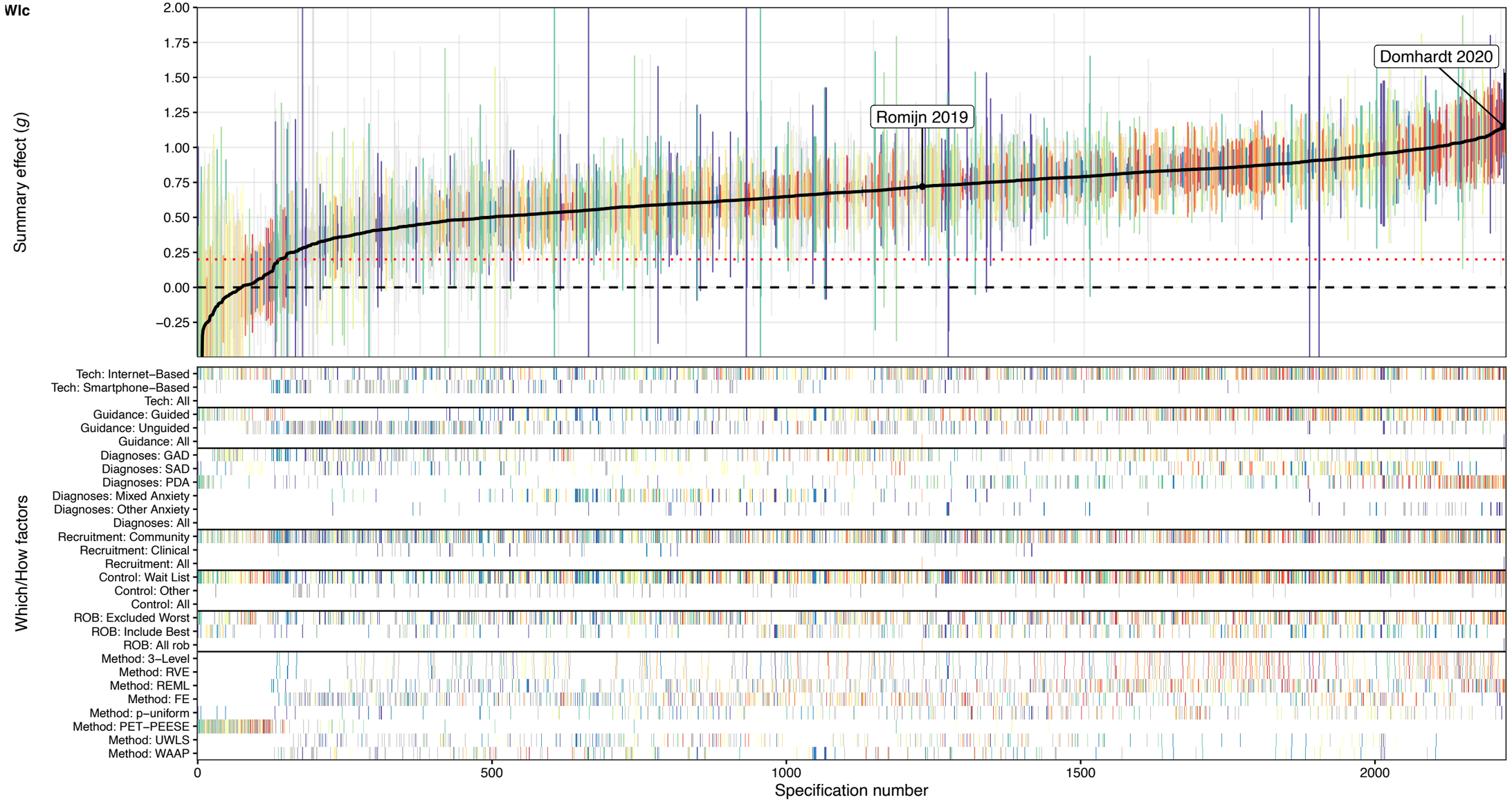
level meta-analytical models, RVE = robust variance estimation, REML = restricted maximum likelihood estimation, FE = fixed effect model, *p*-uniform*, PET-PEESE, UWLS, and WAAP). Each vertical column is color-coded, signifying the number of samples included in a specification (hot spectral colors for more included samples vs. cool spectral colors for less included samples). The overall pattern of the specification curve indicates that barely any meta-analyses could exist on clinical samples.

Figure A.10. ***Raincloud plot of all meta-analyses on digital interventions for anxiety, grouped by different control groups***



Note. Raincloud plots consist of three parts and depict the distribution of data (the cloud), a box-plot, and the raw data (the rain). They depict and visualize the distribution of summary effect sizes from all possible meta-analyses (produced by the multiverse meta-analysis) having either a waitlist control group or other (active control) group.

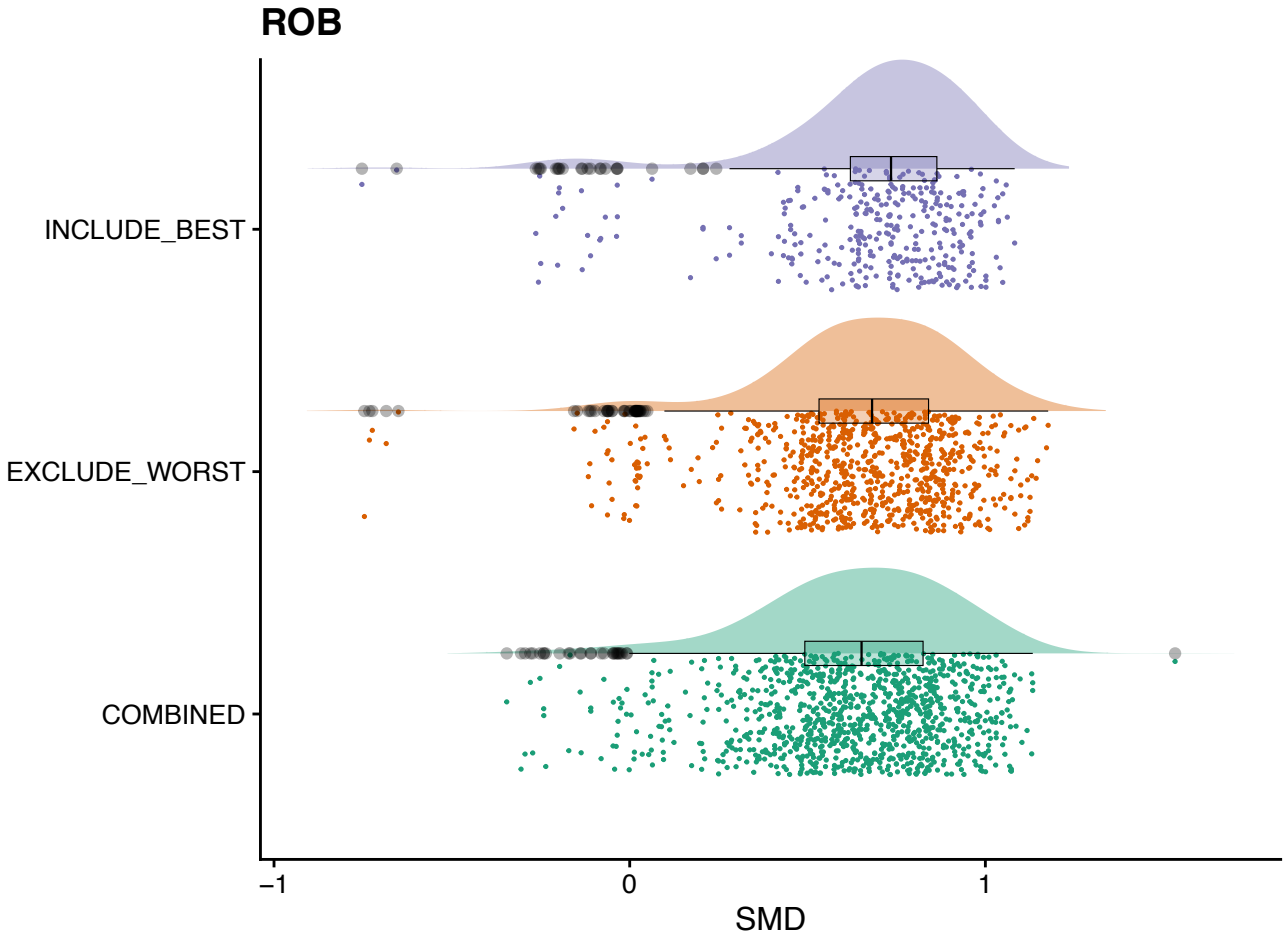
Figure A.11. ***Descriptive specification curve highlighting only meta-analyses that used waitlist control groups as a comparsion***



Note. The top panel shows the meta-analytic summary effects (*g*) for each specification with their 95% confidence interval. The summary effects are sorted by their magnitude, ranging from lower to higher. Connecting the different summary effects results in the solid line, which is the specification curve. A horizontal dashed line of no effect is shown at $g = 0$ and a red dotted line is shown to indicate a small, yet clinically

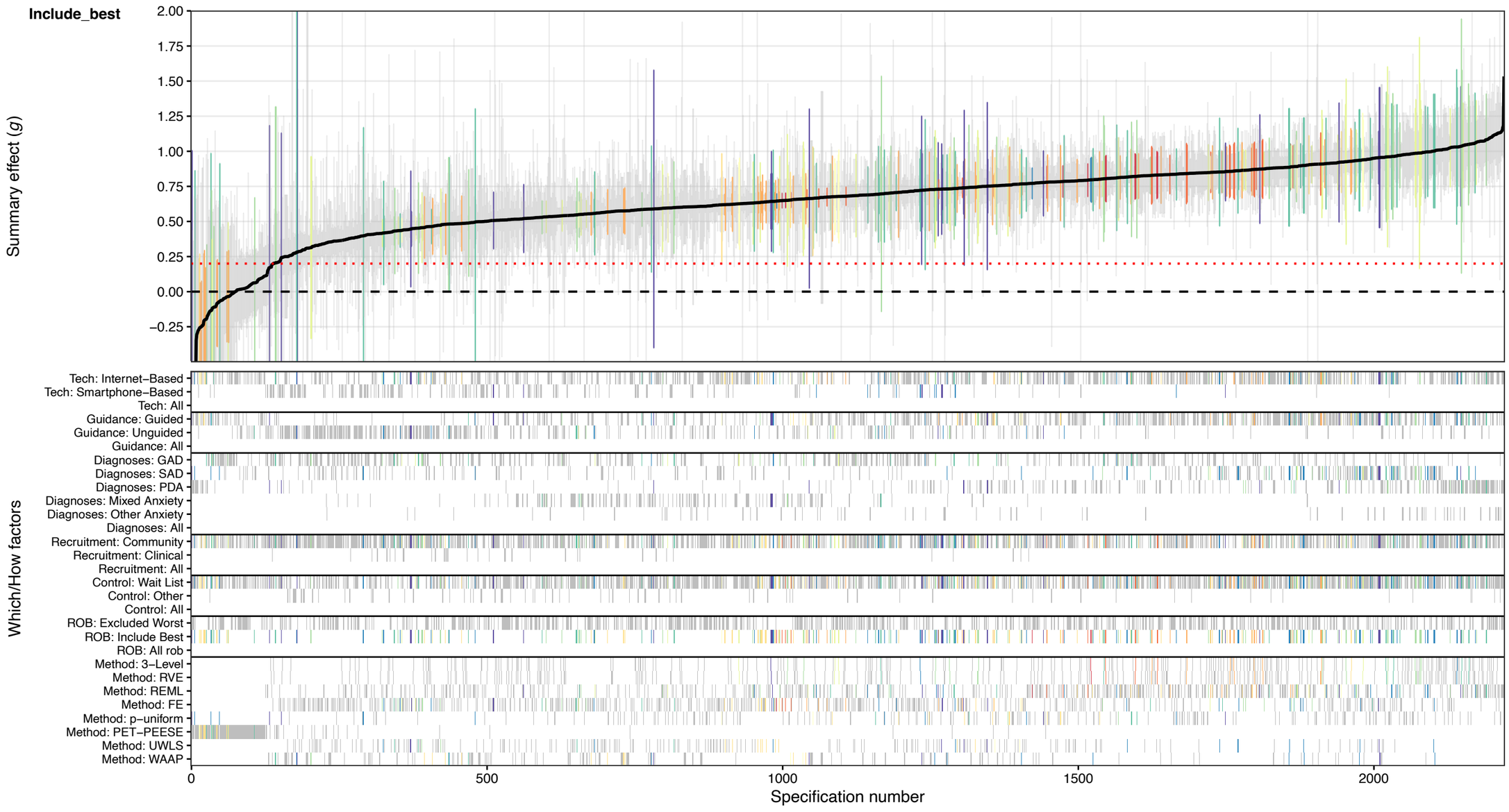
relevant effect size at $g = 0.3$. Published meta-analyses are shown on their respective position on the specification curve. The vertical columns in the bottom panel represent factor combinations of all *Which* factors. These include different technologies, types of guidance, diagnoses, recruitment strategies, control conditions, and strategies to deal with risk of biases. The *How* factors include several meta-analytical estimators: 3-level meta-analytical models, RVE = robust variance estimation, REML = restricted maximum likelihood estimation, FE = fixed effect model, *p*-uniform*, PET-PEESE, UWLS, and WAAP). Each vertical column is color-coded, signifying the number of samples included in a specification (hot spectral colors for more included samples vs. cool spectral colors for less included samples). The overall pattern of the specification curve indicates that waitlist control groups produce much larger effect sizes than meta-analyses comparing an intervention to another type of control condition.

Figure A.12. ***Raincloud plot of all meta-analyses on digital interventions for anxiety, grouped by different strategies to handle high risk of bias studies***



Note. Raincloud plots consist of three parts and depict the distribution of data (the cloud), a box-plot, and the raw data (the rain). They depict and visualize the distribution of summary effect sizes from all possible meta-analyses (produced by the multiverse meta-analysis) having either excluded high risk of bias studies, only included low risk of bias studies, or included all studies.

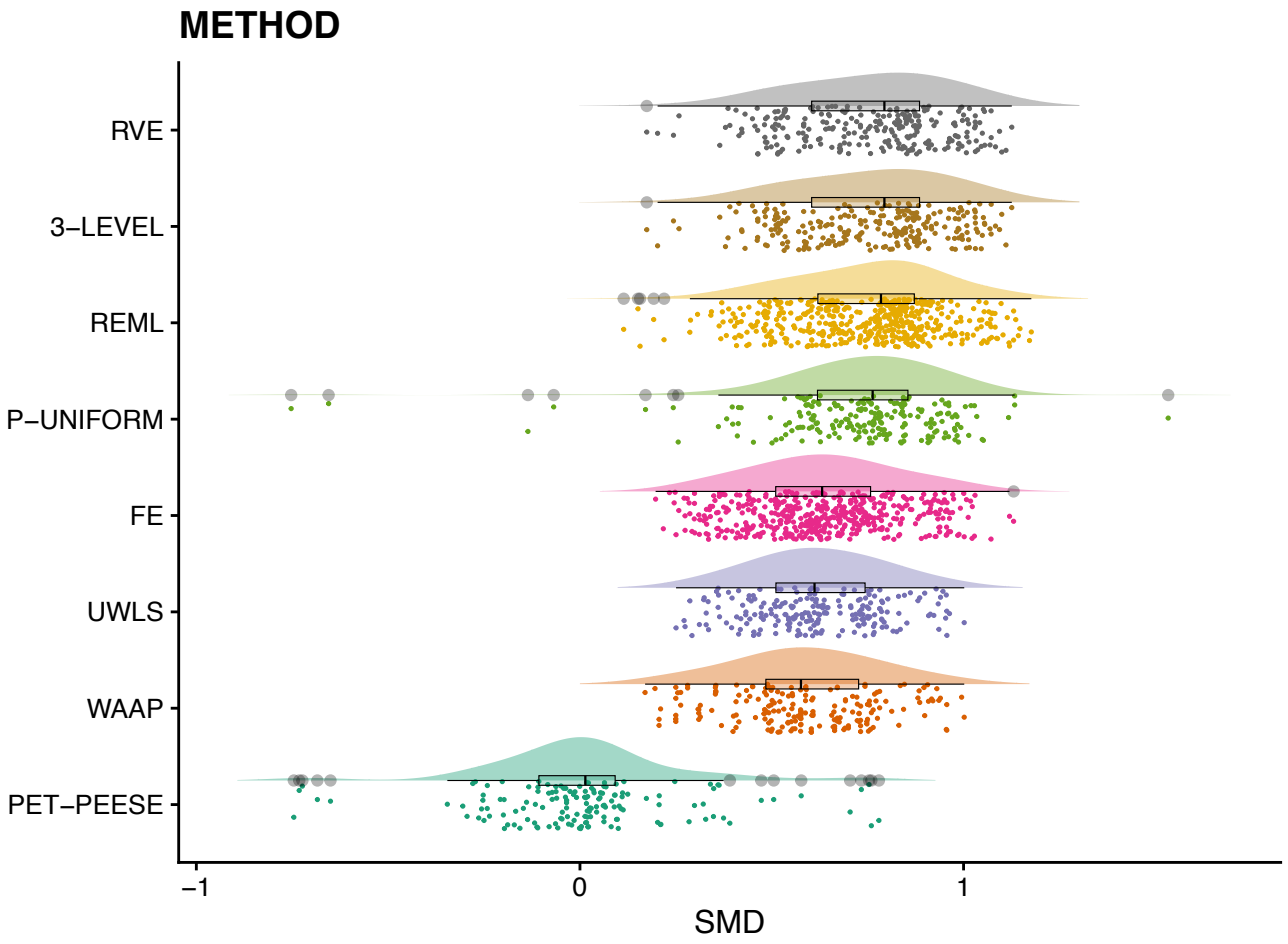
Figure A.13. ***Descriptive specification curve highlighting meta-analyses that only included low risk of bias studies***



Note. The top panel shows the meta-analytic summary effects (*g*) for each specification with their 95% confidence interval. The summary effects are sorted by their magnitude, ranging from lower to higher. Connecting the different summary effects results in the solid line, which is the specification curve. A horizontal dashed line of no effect is shown at $g = 0$ and a red dotted line is shown to indicate a small, yet clinically relevant effect size at $g = 0.3$. Published meta-analyses are shown on their respective position on the specification curve. The vertical columns in the bottom panel represent factor combinations of all *Which* factors. These include different technologies, types of guidance, diagnoses, recruitment strategies, control conditions, and strategies to deal with risk of biases. The *How* factors include several meta-analytical estimators: 3-

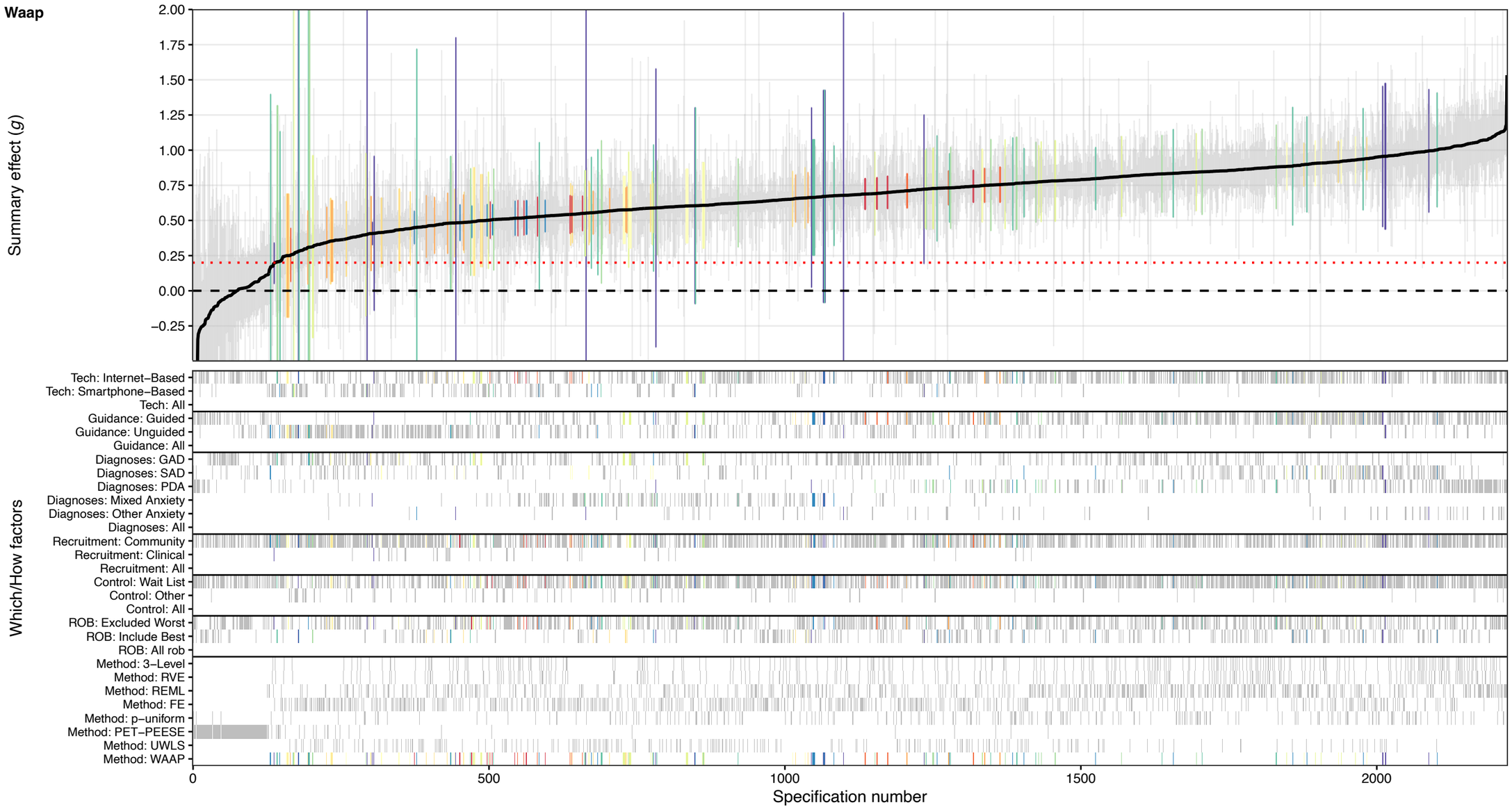
level meta-analytical models, RVE = robust variance estimation, REML = restricted maximum likelihood estimation, FE = fixed effect model, *p*-uniform*, PET-PEESE, UWLS, and WAAP). Each vertical column is color-coded, signifying the number of samples included in a specification (hot spectral colors for more included samples vs. cool spectral colors for less included samples). The overall pattern of the specification curve indicates that including only the best studies (with the lowest risk of bias) produce similar effect sizes than including all studies.

Figure A.14. ***Raincloud plot of all meta-analyses on digital interventions for anxiety, grouped by different meta-analytical estimation methods***



Note. Raincloud plots consist of three parts and depict the distribution of data (the cloud), a box-plot, and the raw data (the rain). They depict and visualize the distribution of summary effect sizes from all possible meta-analyses (produced by the multiverse meta-analysis) calculated with different meta-analytical methods (p-uniform, RVE, 3-level, WAAP, REML, FE, UWLS, PEt-PEESE).

Figure A.15. ***Descriptive specification curve highlighting meta-analyses that were analyzed using the WAAP method***



Note. The top panel shows the meta-analytic summary effects (*g*) for each specification with their 95% confidence interval. The summary effects are sorted by their magnitude, ranging from lower to higher. Connecting the different summary effects results in the solid line, which is the specification curve. A horizontal dashed line of no effect is shown at $g = 0$ and a red dotted line is shown to indicate a small, yet clinically relevant effect size at $g = 0.3$. Published meta-analyses are shown on their respective position on the specification curve. The vertical columns in the bottom panel represent factor combinations of all *Which* factors. These include different technologies, types of guidance, diagnoses, recruitment strategies, control conditions, and strategies to deal with risk of biases. The *How* factors include several meta-analytical estimators: 3-

level meta-analytical models, RVE = robust variance estimation, REML = restricted maximum likelihood estimation, FE = fixed effect model, *p*-uniform*, PET-PEESE, UWLS, and WAAP). Each vertical column is color-coded, signifying the number of samples included in a specification (hot spectral colors for more included samples vs. cool spectral colors for less included samples). The overall pattern of the specification curve indicates that using WAAP produces similar results as other methods.

References

Adelman, C. B., Panza, K. E., Bartley, C. A., Bontempo, A., & Bloch, M. H. (2014). A Meta-Analysis of Computerized Cognitive-Behavioral Therapy for the Treatment of DSM-5 Anxiety Disorders. *The Journal of Clinical Psychiatry*, *75*(7), 11699. https://doi.org/10.4088/JCP.13r08894

Andersson, G., Carlbring, P., Titov, N., & Lindefors, N. (2019). Internet Interventions for Adults with Anxiety and Mood Disorders: A Narrative Umbrella Review of Recent Meta-Analyses. *The Canadian Journal of Psychiatry*, *64*(7), 465–470. https://doi.org/10.1177/0706743719839381

Andrews, G., Basu, A., Cuijpers, P., Craske, M. G., McEvoy, P., English, C. L., & Newby, J. M. (2018). Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: An updated meta-analysis. *Journal of Anxiety Disorders*, *55*, 70–78. https://doi.org/10.1016/j.janxdis.2018.01.001

Apolinário-Hagen, J. (2019). Internet-Delivered Psychological Treatment Options for Panic Disorder: A Review on Their Efficacy and Acceptability. *Psychiatry Investigation*, *16*(1), 37–49. https://doi.org/10.30773/pi.2018.06.26

Bartoš, F., Maier, M., Quintana, D. S., & Wagenmakers, E.-J. (2022). Adjusting for Publication Bias in JASP and R: Selection Models, PET-PEESE, and Robust Bayesian Meta-Analysis. *Advances in Methods and Practices in Psychological Science*, *5*(3), 251524592211092. https://doi.org/10.1177/25152459221109259

Boutron, I., Moher, D., Tugwell, P., Giraudeau, B., Poiraudeau, S., Nizard, R., & Ravaud, P. (2005). A checklist to evaluate a report of a nonpharmacological trial (CLEAR NPT) was developed using consensus. *Journal of Clinical Epidemiology*, *58*(12), 1233–1240. https://doi.org/10.1016/j.jclinepi.2005.05.004

Chisholm, D., Sweeny, K., Sheehan, P., Rasmussen, B., Smit, F., Cuijpers, P., & Saxena, S. (2016). Scaling-up treatment of depression and anxiety: A global return on investment analysis. *The Lancet Psychiatry*, *3*(5), 415–424. https://doi.org/10.1016/S2215-0366(16)30024-4

Cuijpers, P., Turner, E. H., Koole, S. L., van Dijke, A., & Smit, F. (2014). What is the threshold for a clinically relevant effect? The case of major depressive disorders. *Depression and Anxiety*, *31*(5), 374–378. https://doi.org/10.1002/da.22249

Domhardt, M., Letsch, J., Kybelka, J., Koenigbauer, J., Doebler, P., & Baumeister, H. (2020). Are Internet- and mobile-based interventions effective in adults with diagnosed panic disorder and/or agoraphobia? A systematic review and meta-analysis. *Journal of Affective Disorders*, *276*, 169–182. https://doi.org/10.1016/j.jad.2020.06.059

Eilert, N., Enrique, A., Wogan, R., Mooney, O., Timulak, L., & Richards, D. (2021). The effectiveness of Internet-delivered treatment for generalized anxiety disorder: An updated systematic review and meta-analysis. *Depression and Anxiety*, *38*(2), 196–219. https://doi.org/10.1002/da.23115

Eysenck, H. j. (1995). Meta-analysis or best-evidence synthesis? *Journal of Evaluation in Clinical Practice*, *1*(1), 29–36. https://doi.org/10.1111/j.1365-2753.1995.tb00005.x

Firth, J., Torous, J., Nicholas, J., Carney, R., Rosenbaum, S., & Sarris, J. (2017). Can smartphone mental health interventions reduce symptoms of anxiety? A meta-analysis of randomized controlled trials. *Journal of Affective Disorders*, *218*, 15–22. https://doi.org/10.1016/j.jad.2017.04.046

Fusar-Poli, P., & Radua, J. (2018). Ten simple rules for conducting umbrella reviews. *Evidence Based Mental Health*, *21*(3), 95–100. https://doi.org/10.1136/ebmental-2018-300014

Gál, É., Ștefan, S., & Cristea, I. A. (2021). The efficacy of mindfulness meditation apps in enhancing users' well-being and mental health related outcomes: A meta-analysis of randomized controlled trials. *Journal of Affective Disorders*, *279*, 131–142. https://doi.org/10.1016/j.jad.2020.09.134

Goldberg, S. B., Lam, S. U., Simonsson, O., Torous, J., & Sun, S. (2022). Mobile phone-based interventions for mental health: A systematic meta-review of 14 meta-analyses of randomized controlled trials. *PLOS Digital Health*, *1*(1), e0000002. https://doi.org/10.1371/journal.pdig.0000002

Gosling, C. J., Solanes, A., Fusar-Poli, P., & Radua, J. (2023). metaumbrella: The first comprehensive suite to perform data analysis in umbrella reviews with stratification of the evidence. *BMJ Mental Health*, *26*(1), e300534. https://doi.org/10.1136/bmjment-2022-300534

Guo, S., Deng, W., Wang, H., Liu, J., Liu, X., Yang, X., He, C., Zhang, Q., Liu, B., Dong, X., Yang, Z., Li, Z., & Li, X. (2021). The efficacy of internet-based cognitive behavioural therapy for social anxiety disorder: A systematic review and meta-analysis. *Clinical Psychology & Psychotherapy*, *28*(3), 656–668. https://doi.org/10.1002/cpp.2528

Higgins, J. P. T., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., Savović, J., Schulz, K. F., Weeks, L., & Sterne, J. A. C. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, *343*, d5928. https://doi.org/10.1136/bmj.d5928

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

Ioannidis, J. P. A. (2009). Integration of evidence from multiple meta-analyses: A primer on umbrella reviews, treatment networks and multiple treatments meta-analyses.

Canadian Medical Association Journal, *181*(8), 488–493. https://doi.org/10.1503/cmaj.081086

Ioannidis, J. P. A. (2017). Next-generation systematic reviews: Prospective meta-analysis, individual-level data, networks and umbrella reviews. *British Journal of Sports Medicine*, *51*(20), 1456–1458. https://doi.org/10.1136/bjsports-2017-097621

Kampmann, I. L., Emmelkamp, P. M. G., & Morina, N. (2016). Meta-analysis of technology-assisted interventions for social anxiety disorder. *Journal of Anxiety Disorders*, *42*, 71–84. https://doi.org/10.1016/j.janxdis.2016.06.007

Lecomte, T., Potvin, S., Corbière, M., Guay, S., Samson, C., Cloutier, B., Francoeur, A., Pennou, A., & Khazaal, Y. (2020). Mobile Apps for Mental Health Issues: Meta-Review of Meta-Analyses. *JMIR MHealth and UHealth*, *8*(5), e17458. https://doi.org/10.2196/17458

Linardon, J., Cuijpers, P., Carlbring, P., Messer, M., & Fuller-Tyszkiewicz, M. (2019). The efficacy of app-supported smartphone interventions for mental health problems: A meta-analysis of randomized controlled trials. *World Psychiatry*, *18*(3), 325–336. https://doi.org/10.1002/wps.20673

Loo Gee, B., Griffiths, K. M., & Gulliver, A. (2016). Effectiveness of mobile technologies delivering Ecological Momentary Interventions for stress and anxiety: A systematic review. *Journal of the American Medical Informatics Association*, *23*(1), 221–229. https://doi.org/10.1093/jamia/ocv043

Mor, S., Grimaldos, J., Tur, C., Miguel, C., Cuijpers, P., Botella, C., & Quero, S. (2021). Internet- and mobile-based interventions for the treatment of specific phobia: A systematic review and preliminary meta-analysis. *Internet Interventions*, *26*, 100462. https://doi.org/10.1016/j.invent.2021.100462

NHLBI, N. (2020). *Study quality assessment tools*. National Heart, Lung and Blood Institute.

Olthuis, J. V., Watt, M. C., Bailey, K., Hayden, J. A., & Stewart, S. H. (2016). Therapist-supported Internet cognitive behavioural therapy for anxiety disorders in adults. *Cochrane Database of Systematic Reviews*, *3*. https://doi.org/10.1002/14651858.CD011565.pub2

Papatheodorou, S. (2019). Umbrella reviews: What they are and why we need them. *European Journal of Epidemiology*, *34*(6), 543–546. https://doi.org/10.1007/s10654-019-00505-6

Păsărelu, C. R., Andersson, G., Bergman Nordgren, L., & Dobrean, A. (2017). Internet-delivered transdiagnostic and tailored cognitive behavioral therapy for anxiety and depression: A systematic review and meta-analysis of randomized controlled trials. *Cognitive Behaviour Therapy*, *46*(1), 1–28.

Patel, V., Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., Chisholm, D., Collins, P. Y., Cooper, J. L., Eaton, J., Herrman, H., Herzallah, M. M., Huang, Y., Jordans, M. J. D., Kleinman, A., Medina-Mora, M. E., Morgan, E., Niaz, U., Omigbodun, O., … UnÜtzer, Jü. (2018). The Lancet Commission on global mental health and sustainable development. *The Lancet*, *392*(10157), 1553–1598. https://doi.org/10.1016/S0140-6736(18)31612-X

Pauley, D., Cuijpers, P., Papola, D., Miguel, C., & Karyotaki, E. (2021). Two decades of digital interventions for anxiety disorders: A systematic review and meta-analysis of treatment effectiveness. *Psychological Medicine*, 1–13. https://doi.org/10.1017/S0033291721001999

Plessen, C. Y., Karyotaki, E., Miguel, C., Ciharova, M., & Cuijpers, P. (2023). Exploring the efficacy of psychotherapies for depression: A multiverse meta-analysis. *BMJ Ment Health*, *26*(1). https://doi.org/10.1136/bmjment-2022-300626

Richards, D., Richardson, T., Timulak, L., & McElvaney, J. (2015). The efficacy of internet-delivered treatment for generalized anxiety disorder: A systematic review and meta-analysis. *Internet Interventions*, *2*(3), 272–282. https://doi.org/10.1016/j.invent.2015.07.003

Romijn, G., Batelaan, N., Kok, R., Koning, J., Balkom, A. van, Titov, N., & Riper, H. (2019). Internet-Delivered Cognitive Behavioral Therapy for Anxiety Disorders in Open Community Versus Clinical Service Recruitment: Meta-Analysis. *Journal of Medical Internet Research*, *21*(4), e11706. https://doi.org/10.2196/11706

Sharpe, D., & Poets, S. (2020). Meta-analysis as a response to the replication crisis. *Canadian Psychology/Psychologie Canadienne*, *61*(4), 377–387. https://doi.org/10.1037/cap0000215

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., Moher, D., Tugwell, P., Welch, V., Kristjansson, E., & Henry, D. A. (2017). AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ*, j4008. https://doi.org/10.1136/bmj.j4008

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 Word Solution. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2160588

Slavin, R. E. (1995). Best evidence synthesis: An intelligent alternative to meta-analysis. *Journal of Clinical Epidemiology*, *48*(1), 9–18. https://doi.org/10.1016/0895-4356(94)00097-a

Solmi, M., Correll, C. U., Carvalho, A. F., & Ioannidis, J. P. A. (2018). The role of meta-analyses and umbrella reviews in assessing the harms of psychotropic medications: Beyond qualitative synthesis. *Epidemiology and Psychiatric Sciences*, *27*(6), 537–542. https://doi.org/10.1017/S204579601800032X

Stanley, T. D. (2017). Limitations of PET-PEESE and Other Meta-Analysis Methods. *Social Psychological and Personality Science*, *8*(5), 581–591. https://doi.org/10.1177/1948550617693062

Stanley, T. D., Doucouliagos, H., & Ioannidis, J. P. A. (2022). Beyond Random Effects: When Small-Study Findings Are More Heterogeneous. *Advances in Methods and Practices in Psychological Science*, *5*(4), 25152459221120428. https://doi.org/10.1177/25152459221120427

Stech, E. P., Lim, J., Upton, E. L., & Newby, J. M. (2020). Internet-delivered cognitive behavioral therapy for panic disorder with or without agoraphobia: A systematic review and meta-analysis. *Cognitive Behaviour Therapy*, *49*(4), 270–293.

Sterne, J. A. C., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A.-W., Churchill, R., Deeks, J. J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y. K., Pigott, T. D., … Higgins, J. P. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*, *355*, i4919. https://doi.org/10.1136/bmj.i4919

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H.-Y., Corbett, M. S., Eldridge, S. M., Emberson, J. R., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., … Higgins, J. P. T. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ*, l4898. https://doi.org/10.1136/bmj.l4898

Stratton, E., Lampit, A., Choi, I., Calvo, R. A., Harvey, S. B., & Glozier, N. (2017). Effectiveness of eHealth interventions for reducing mental health conditions in employees: A systematic review and meta-analysis. *PLOS ONE*, *12*(12), e0189904. https://doi.org/10.1371/journal.pone.0189904

van Aert, R. C. M., & van Assen, M. A. L. M. (2020). P-uniform* [Preprint]. In *MetaArXiv*. https://doi.org/10.31222/osf.io/zqjr9

Versluis, A., Verkuil, B., Spinhoven, P., van der Ploeg, M. M., & Brosschot, J. F. (2016). Changing Mental Health and Positive Psychological Well-Being Using Ecological Momentary Interventions: A Systematic Review and Meta-analysis. *Journal of Medical Internet Research*, *18*(6), e152. https://doi.org/10.2196/jmir.5642

Viechtbauer, W. (2010). Conducting Meta-Analyses in *R* with the **metafor** Package. *Journal of Statistical Software*, *36*(3). https://doi.org/10.18637/jss.v036.i03

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

Weisel, K. K., Fuhrmann, L. M., Berking, M., Baumeister, H., Cuijpers, P., & Ebert, D. D. (2019). Standalone smartphone apps for mental health—A systematic review and meta-analysis. *Npj Digital Medicine*, *2*(1), 118. https://doi.org/10.1038/s41746-019-0188-8

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18