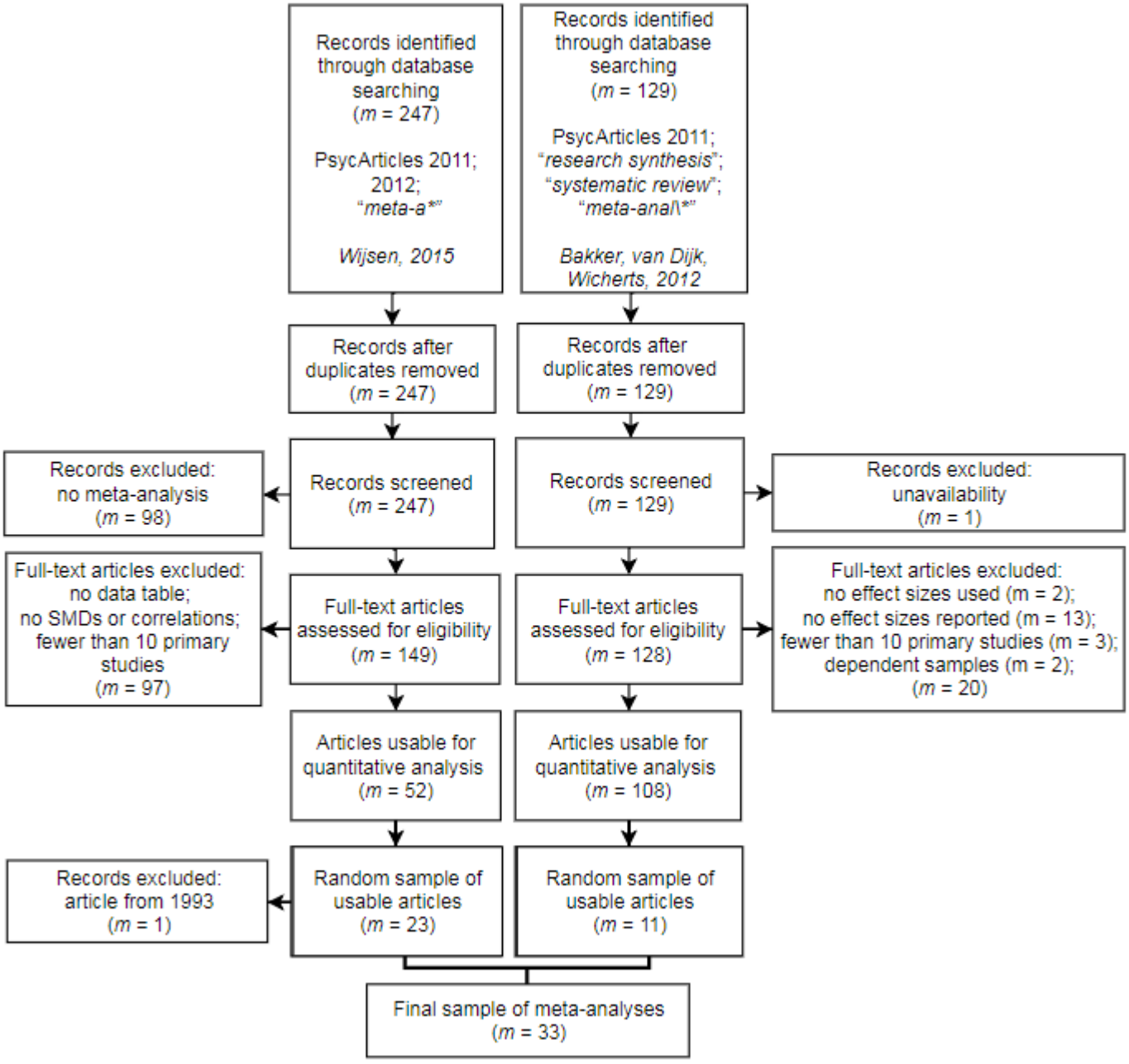
Sampling scheme and flowcharts

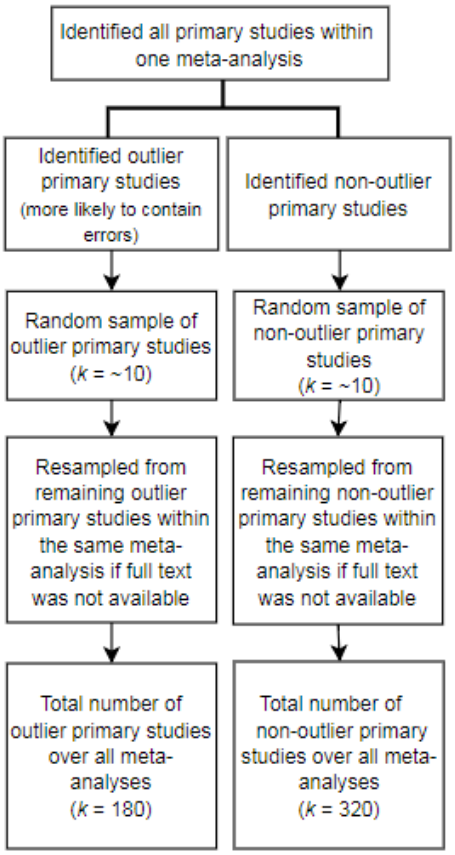
Meta-analysis selection



Our sample of meta-analyses (total $m = 33$) was chosen from previous research that had coded materials readily available (Bakker, van Dijk, and Wicherts, 2012; Wijsen, 2015). Suitable meta-analyses from the first study (Wijsen, 2015) were published in 2011 or 2012 and available in the database PsycArticles with "*meta-a**" as a search criterion. The goal of the meta-analysis selection was to obtain a representative sample of psychological studies, but the inclusion criteria may have affected the representativeness of the sample. More specifically, a large number ($m = 97$) of meta-analyses could not be included because they omitted necessary basic statistics. For instance, they did not report a data table with primary studies, effect sizes and standard errors or within-study variances, the meta-analyses did not contain the appropriate effect sizes (i.e., SMDs or correlations), or did not meet the minimum requirement of ten primary studies, to ensure power of the meta-analysis was large enough to provide accurate estimation of effect sizes. The number of eligible meta-analyses remaining was 52, of which 23 were randomly sampled by (Wijsen, 2015) because of feasibility constraints.

The final 10 meta-analyses (Bakker et al., 2012) were selected from the total number of eligible peer-reviewed articles published in 2011 from PsycArticles (m= 129), containing the strings "*research synthesis*", "*systematic review*", or "*meta-anal**" in the title and/or abstract. A total of 21 meta-analyses were excluded because the meta-analyses did not contain (m=2) or did not report (m=13) effect sizes of primary studies, there were fewer than 10 primary studies (m=3), unavailability (m=1), or samples were dependent, leading to biased standard errors (m=2). Of the remaining 108 meta-analyses, Bakker et al. randomly selected 10 (10% of the total). In 16 cases, there were multiple meta-analyses within one article, in which case one meta-analysis was randomly selected. The total number of selected meta-analytic articles was 33, which contained 1951 primary effect sizes, of which we sampled 500 to reproduce.

Primary study selection



Our selection of primary studies is displayed in the flowchart above. When selecting primary studies from the meta-analyses, we wanted to sample both primary studies that could be considered outliers compared to the rest of the primary studies in the meta-analysis, as well as non-outlier primary studies. Some data errors in meta-analyses are likely to result in outliers, which inflate variance estimates (Schmidt and Hunter, 2015). Because of this, we oversampled outlier primary studies to investigate whether outliers were more likely to be erroneous compared to non-outlier primary studies. First, for each of the 33 meta-analyses, we fitted a random-effects model in *R* version 3.5.1 with the metafor package (R Core Team, 2018; Viechtbauer, 2010), after which the leave-one-out function repeatedly fitted the model again, leaving out one study at a time. Outliers were detected by calculating the Q-statistic that is used to test the null hypothesis of homogeneity. The Q-statistic of the complete meta-analysis was compared to the Q-statistic of the complete meta-analysis minus one study. If effect sizes are homogeneous, Q follows a χ^2 distribution with k-1 degrees of freedom, where k is the number of studies. Since the difference be-

tween the two random-effects models is always one study, $k = 1$. If the observed difference between the two Q values exceeded the 95^{th} percentile of the central χ^2 distribution, the left out primary study was classified as an outlier.

For all meta-analyses, we classified all primary studies as either being an outlier or non-outlier. Per meta-analysis, we first took a random sample from the collection of outlier primary studies. To ensure a fair distribution of outlier and non-outlier primary studies, we sampled 10 outlier primary studies per meta-analysis [1]. If a meta-analysis contained fewer than 10 outliers, we included all outliers. If we could not retrieve an outlier primary study due to article unavailability, we randomly selected another outlier from the same meta-analysis until none were left. In total, we included 197 primary studies that were considered outliers. After this selection, a total of 303 primary studies remained to be sampled from the non-outlier primary studies within the 33 meta-analyses. The remaining primary studies were evenly selected from the meta-analyses, meaning we randomly chose approximately 10 studies per meta-analysis. If we failed to find the article for a non-outlier primary study, we randomly selected another non-outlier primary study from the same meta-analysis. If the number of primary studies to extract exceeded the total number of primary studies in the meta-analysis, we randomly divided the remaining number of primary studies among the other meta-analyses that had studies left to be sampled.

References

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

R Core Team. (2018). R: A language and environment for statistical computing. Retrieved from https://www.R-project.org

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings*. (pp.64). doi:10.4135/9781483398105

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

Wijsen, L. (2015). *Do psychological effects fade away? meta-analyzing the decline effect.* (Unpublished research report). University of Amsterdam, Amsterdam, the Netherlands.

[1] Inadvertently, in three cases, more than ten outlier primary study effect sizes were sampled; meta-analysis no. 7 and 19 have 13 non-outlier primary studies, meta-analysis no. 9 has 12 non-outlier primary studies.