The effects of educational video-games on students motivation to math: A meta-analysis in K-12

List authors

List affiliations

Analysis Report

Edited by

Claudio Zandonella Callegher*

*e-mail address

claudiozandonella@gmail.com

January 23, 2020

Contents

List of Figures

List of Tables

1 Introduction

In this report the statistical analyses of the article *"The effects of educational video-games on students motivation to math: A meta-analysis in K-12"* are presented. The aim of the meta-analysis was to synthesized results of the studies concerning the impact of educational video-games on students' motivation towards mathematics.

Here, we will focus only on the statistical analysis. For theoretical aspects, study selection, and results interpretation the reader can refer directly to the article.

1.1 Report sections

The analysis report is divided into different sections:

- **Section 2:** the statistical approach and the plan of analysis are presented.
- **Section 3:** the dataset is presented with a brief description of each variable and prepared for the analysis.
- **Section 4:** the descriptive statistics are presented.
- **Section 5:** the results of the meta-analysis are described.

2 Statistical Approach

In this section, the statistical approach and the plan of the meta-analysis are presented. Fist, we describe the measure of effect size used to evaluate effectiveness of the interventions. Subsequently, we discuss the reasons why multilevel meta-analysis is used to account for the dependence of multiple effect sizes within the same studies. Finally, the plan of the meta-analysis is summarized.

2.1 Measure of effect size

2.1.1 The pre- post- control group design

Selected studies were characterized by a pre- post- control group design (PPC). In a PPC design, participants are randomly assigned to one (or more) experimental group and to one (or more) control group. Both groups are evaluated before (pre-test score) and after (post-test score) the experimental group is exposed to a treatment.

The PPC design allows to evaluate the efficacy of the treatment taking into account pre-existing differences between the two groups and concurrent factors or events other than the treatment that produce changes in the outcome variable.

The efficacy of the treatment can evaluated considering the standardized mean change in both groups. The standardized mean change of each group is defined as the mean difference between post-test (μ_{post}) and pre-test (μ_{pre}) scores, divided by the standard deviation (σ):

$$\delta = \frac{\mu_{post} - \mu_{pre}}{\sigma}. \tag{1}$$

Thus, assuming common standard deviation (σ), the efficacy of the treatment can be defined as the difference in standardized mean change between the experimental group and the control group:

$$\Delta = \delta_{Eg} - \delta_{Cg} = \frac{(\mu_{Eg,post} - \mu_{Eg,pre}) - (\mu_{Cg,post} - \mu_{Cg,pre})}{\sigma}, \tag{2}$$

where Eg stands for experimental group and Cg stands for control group.

2.1.2 The effect size estimate

To estimate the effect size, there are three alternatives that differ in the way the common standard deviation (σ) is estimated. Morris (2008) discuss and evaluate the three alternatives:

1. $\boldsymbol{d_{ppc1}}$: the effect size is computed using separate estimates of pre-test standard deviation for the experimental group (Eg) and the control group (Cg). That is,

$$d_{ppc1} = c_{Eg}\frac{(M_{Eg,post} - M_{Eg,pre})}{SD_{Eg,pre}} - c_{Cg}\frac{(M_{Cg,post} - M_{Cg,pre})}{SD_{Cg,pre}}, \tag{3}$$

   where c_{Eg} and c_{Cg} are bias adjustments for small samples, M_{pre} and M_{post} are the mean scores in the pre-test and post-test, and SD_{pre} is the standard deviation in the pre test.

2. $\boldsymbol{d_{ppc2}}$: the effect size is computed by pooling the data from the experimental and control group in the pre-test to estimate the population standard deviation. That is,

$$d_{ppc2} = c_P \left[\frac{(M_{Eg,post} - M_{Eg,pre}) - (M_{Cg,post} - M_{Cg,pre})}{SD_{pre}} \right], \tag{4}$$

where SD_{pre}[1] is the pooled standard deviation in the pre-test scores, and c_P[2] is a bias adjustment for small sample size.

3. $\boldsymbol{d_{ppc3}}$: the effect size is computed by pooling the data from the experimental and control group in both the pre-test and post-test to estimate the population standard deviation.That is,

$$d_{ppc3} = c_{PP} \left[\frac{(M_{Eg,post} - M_{Eg,pre}) - (M_{Cg,post} - M_{Cg,pre})}{SD_{pre+post}} \right], \tag{5}$$

where $SD_{pre+post}$[3] is the pooled standard deviation in the pre-test and post-test scores, and c_{PP}[4] is a bias adjustment for small sample size.

Morris (2008) suggested the d_{ppc2} as the favourite effect size estimate in terms of bias, precision, and robustness to heterogeneity of variance. Given the assumption of homogeneity of variance in the two populations, d_{ppc2} allows a better estimate of the population standard deviation by pooling the data from the experimental and control groups in the pre-test. On the contrary, d_{ppc1} uses separate estimates of the sample standard deviation for experimental and control groups and d_{ppc3} includes in the pooled standard deviation also results from the post-test of the two groups. These solutions are not optimal as, in the first case, we do not take advantage of the assumption of homogeneity of variance, and in the second case post-test variances tend to be larger than pre-test variances given possible interaction between treatment effect and individual differences.

The formula to compute the variance of d_{ppc2} is provided by Morris (2008, see p.373 eq.25):

$$\sigma^2(d_{ppc2}) = 2(c_p^2)(1-\rho)\left(\frac{n_{Eg}+n_{Cg}}{n_{Eg}n_{Cg}}\right)\left(\frac{n_{Eg}+n_{Cg}-2}{n_{Eg}+n_{Cg}-4}\right)\left(1+\frac{d_{ppc2}^2}{2(1-\rho)\left(\frac{n_{Eg}+n_{Cg}}{n_{Eg}n_{Cg}}\right)}\right) - d_{ppc2}^2, \tag{6}$$

where ρ is the the correlation between pre- and post-test scores, and n_{Eg} and n_{Cg} are the two groups sample size.

Following recommendations of the author, we used the d_{ppc2} to estimate the effect size of the studies included in the meta-analysis.

2.2 Multilevel meta-analysis

Traditional meta-analysis approaches are based on the assumption that each effect sizes is independent, so each study should contribute with only one effect size. However, dependency among effect sizes is very

[1] $SD_{pre} = \sqrt{\frac{(n_{Eg}-1)SD_{Eg,pre}^2+(n_{Cg}-1)SD_{Cg,pre}^2}{n_{Eg}+n_{Cg}-2}}$

[2] $c_P = 1 - \frac{3}{4(n_{Eg}+n_{Cg}-2)-1}$

[3] $SD_{pre+post} = \sqrt{\frac{(n_{Eg}-1)SD_{Eg,pre}^2+(n_{Cg}-1)SD_{Cg,pre}^2+(n_{Eg}-1)SD_{Eg,post}^2+(n_{Cg}-1)SD_{Cg,post}^2}{2(n_{Eg}+n_{Cg}-2)}}$

[4] $c_{PP} = 1 - \frac{3}{4(2n_{Eg}+2n_{Cg}-4)-1}$

common. Studies could evaluate the same subjects on multiple outcomes providing measures that are clearly not independent, or, even when outcomes are measured on independent samples, multiple effects within the same study are not independent. In fact, they share other common aspect as the design of the experiment, instruments used, treatment characteristics, research group, geographical region, etc., that may influence the effects to be more similar to each other.

In order to take into account the dependency between effect sizes, different approaches are proposed (Moeyaert et al., 2017; Pigott & Polanin, 2019):

1. **Averaging effect sizes**. When similar measures of the same underlying construct are provided, it is reasonable to use the average of the effect sizes to summarize the study results. Borenstein, Hedges, Higgins, and Rothstein (2009, see Part 5: Complex data structure) describe how to compute a composite effect size and variance taking according to the dependence structures of the effect size within the study.
2. **Robust variance estimation (RVE)**. Hedges, Tipton, and Johnson (2010) recommend to use the RVE when evaluating dependent effects in a meta-analysis. The overall effect size over studies is computed as a weighted mean of the observed effect size and the sampling variance estimate is obtained by means of an approximation of the covariance-matrix of each study. Hedges et al. (2010) demonstrate that RVE meta-analysis with non-independent effect sizes produces valid results even in the case of misspecified covariance structure. As the number of the studies increase the results will converge to the correct values. Moreover, Tipton (2015) provided adjusted estimators to the RVE that increase the reliability of the results even when the number of studies is small.
3. **Multilevel meta-analysis**. All random-effects meta-analysis are multilevel as they assume a two-level structure where the observed effects are an estimation of each study "true" effect that is sampled from an overarching distribution of "true" effects. However, in this case we refer to three-level meta-analysis model, where participants are included at level 1, measured outcomes at level 2, and study at level 3. In this three level approach, the dependency between effect size within studies is automatically accounted for in the covariance matrix (Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2015).

Moeyaert et al. (2017) compared the three approaches and reported that both the RVA and the multilevel approach give unbiased parameter estimates, standard errors and variances. Whereas, averaging effect sizes is too conservative because in general standard errors are overestimated. Moreover, the RVA and the multilevel approach allow to include in the meta-analytic model predictors at the outcome level to account for differences between outcomes within the same studies.

In the present meta-analysis, we decided to adopt a multilevel approach as it allows us to specify the kind of dependency between effect size specifically for each study. Multilevel meta-analysis offers a great flexibility in accommodating different type of dependency among effect sizes within studies.

On the contrary, RVA approach allows only to decide between a correlated effects dependence structure or a hierarchical dependence structure for all the studies (Hedges et al., 2010; Tipton, 2015). In the first case, within each study outcomes are assumed to be measured on the same underlying units. In the second case, multiple (independent) outcomes are measured within studies, but they are related by other aspects such as research organizations, research labs, or research groups.

2.3 Analysis plan

The analysis are conducted with R software (version 3, ., 6.1, ; R Core Team, 2018). First, we load and prepare the dataset for the analysis, computing the effect size estimate and variance for each outcome. As presented in Formula 6, values of the correlation between pre-test and post-test scores are needed to the compute d_{ppc2} variance. However, none of the studies reported this information. To overcome this issue, d_{ppc2} variance is computed with different correlation values and their influence on the meta-analysis results will be subsequently evaluated.

Next, descriptive statistics are presented to evaluate the main characteristics of the data included in the analysis.

Multilevel meta analysis is conduct using the restricted maximum likelihood method with the R-package `metafor` (Viechtbauer, 2010). Variance-covariance matrix is defined for each study according to the specific outcomes dependence structure. However, studies with multiple measure on the same subjects did not reported the correlation between outcomes needed to specify dependence structure. In these cases, variance-covariance matrix is computed with different correlation values and their influence on the meta-analysis results will be subsequently evaluated. Results are presented and heterogeneity between studies is assessed through inspection of forest plot and evaluation of the Q-statistic (Hedges & Olkin, 2014). Under the null hypothesis, the Q-statistic is distributed as a chi-square with degrees of freedom equal to the number of studies minus one. A significant chi-value indicates the presence of heterogeneity across studies. Moreover, to estimate the magnitude of the heterogeneity the I^2 index is reported (i.e., the proportion of observed variance that reflects real and not random difference between studies effect sizes; Borenstein et al., 2009). High values of I^2 suggest that differences between results are related to real differences across studies (i.e., different constructs or different study design). On the contrary, low values of I^2 suggest that results across studies are similar and possible differences are related to random sampling.

Next, we present the alternative results obtained using the different meta-analytic approaches presented above, which are robust variance estimation (RVE) approach and averaging effect sizes approach. RVE is conducted using the R-package `robumeta` (Fisher & Tipton, 2015; Fisher, Tipton, & Zhipeng, 2017) considering a correlated effects dependence structure. Averaged effect sizes are obtained according to Borenstein et al. (2009) indications as implemented in the R-package `MAd` (Del Re & Hoyt, 2014).

To investigate robustness of the results, a sensitivity analysis is conducted to evaluate how the values of the correlation between pre-test and post-test scores and the values of the correlation between dependent

outcomes influence the meta analysis results. Moreover, a *Leave-One-Out* (LOO) sensitivity analysis considering the study grouping level is run to assess the possible presence of influential studies. Substantial changes when a single study is removed would be interpreted as lack of homogeneity and unreliable results (Viechtbauer & Cheung, 2010).

Subsequently, publication bias is assessed. To our knowledge, there are no specific methods designed to correctly evaluate publication bias in a multilevel meta-analysis. Funnel plot is presented as a graphical representation of the symmetry of the effects distribution. However, this is not optimal as studies with multiple effects are overrepresented in the plot respect to studies with only one effect. To overcome this issue, we conduct separately two Egger's regression tests to evaluate if effect sizes are associated to sample size or sampling variance of the estimated effect. In the absence of publication bias, the regression coefficient is expected to be zero (Egger, Smith, Schneider, & Minder, 1997; Lin & Chu, 2018). Moreover, the result of the rank correlation test is reported to examine whether the observed outcomes and the corresponding sampling variances are correlated (Begg & Mazumdar, 1994).

Alternatively, funnel plot with the *trim-and-fill* method (Duval & Tweedie, 2000; Rothstein, Sutton, & Borenstein, 2005) is presented considering the aggregated effects.

Finally, the role of possible moderators is examined using mixed-effects meta-regression models, the moderators are included as a fixed effects and are tested using Wald's chi-square (Viechtbauer, 2010). Considering the reduced number of studies and the unequal distribution among the different levels of the moderators, separate analyses are conducted for each moderator. Results have to be interpreted with caution as moderators are not evaluated together to understand the unique variance explained by each moderator.

2.4 Analysis reproducibility

To guarantee the reproducibility of the results, the whole analysis is structured within an *R-project* named `DMGC_Meta.Rproj` that is possible to download from this repository (Todo: add repository link).

The R-package `drake` (Landau, 2018) is used to manage the analysis workflow and to enhance the readability and transparency of the analysis. To know more about `drake` consider the official Git-hub page (https://github.com/ropensci/drake) or the user manual(https://books.ropensci.org/drake/). Summarizing, using `drake` the code of the analysis is organized into different scripts. The user defines the plan of the analysis where each step in the analysis is defined trough functions. Functions can be appropriately defined to obtain desired targets (i.e., R-output with results of interests) and they are declared in another script. Subsequently, `drake` manages the whole analysis recognizing the dependency structure of the different targets. When any change is made to the code `drake` evaluates the analysis and updates the results. Using functions to define each step of the analysis allows to avoid "*coping and paste*" in the code, it makes debugging easier, and it facilitates the reading of the code.

Moreover, the R-package `renv` (Ushey, 2019) is used to manage the dependencies of the R-packages used

in the analysis. The `renv` package allows to create an isolated, portable, and reproducible environment where the analyses are run. To know more about `renv` consider the official documentation (https://rstudio.github.io/renv/articles/renv.html).

Finally, git version control was used to track the changes during the analysis.

2.4.1 R-project structure

The R-project `DMGC_Meta.Rproj` is organized into different folders. In the folder `Data/`, the raw datasets with the information regarding the studies selected in the literature review are stored.

In the folder `R/`, the R-scripts used in the analysis are stored. Using the `drake` package the analysis is organized into different R-scripts files:

- `Settings.R` contains the setting for the R sessions, including R-packages used.
- `Plan.R` contains the plan of the analysis. Where each target (i.e., R-output with results of interests) is defined through functions.
- `Function.R` contains the main functions used in `Plan.R` to obtain the targets of interest.
- `Auxiliary_functions.R` contains other functions used in the analysis.
- `Analysis.R` is the script used to run the whole analysis.

In the folder `Report_analysis/`, it is possible to find the script used to compile the present report.

2.4.2 Run the Analysis

In order to run the analysis follow these steps:

1. Make sure you have already the `renv` R-package installed in your library. If not, run the command in R or R-studio `install.packages("renv")`
2. Open the R-project `DMGC_Meta` by double-clicking the file `DMGC_Meta.Rproj` you can find in the main directory. A new R-studio session should open and a similar message should appear in the console if `renv` was correctly installed:
   `* Project '~/<your_path>/DMGC_Meta' loaded. [renv <version_number>]`
3. Run the line `renv::restore()`, `renv` will ask the permission to install the R-packages used in the analysis, type `y` and return to confirm.
4. Open the file 'R/Analysis.R' and run each line of the sections "Load", "Check", and "Make".
5. Now you can access the targets with the results using the functions `drake::loadd(<name_target>)` and `drake::readd(<name_target>)`.

3 Data Preparation

In this section, first the raw dataset is presented, then the dataset is prepared for the analysis.

3.1 Data presentation

In the folder `Data/`, you can find the file `Dataset.csv` with all the informations about the selected studies. The dataset includes 43 effects sizes grouped within 19 different studies. Each line of the dataset is an effect and the characteristics of each effect are summarized in 26 variables:

1. `study` - Numerical variable that indicates the id number of the study. This id number is used in the analysis to refer to a specific study. Values range from `1` to `19`.
2. `id` - Numerical variable that indicates the id number of the study used during the selection process of eligible studies.
3. `author` - Character string with the names of the authors.
4. `year` - Numerical variable that indicates the year of publication.[5].
5. `pub` - Character string that indicates if the study is published in peer-reviewed journal (`yes`) or conferece papers and unpublished dissertations (`no`).
6. `grade` - Numerical variable that indicates school grade of the participants. Primary school is indicated with `1` and secondary school is indicated with `2`.
7. `weeks` - Numerical variable that indicates the duration of the treatment in weeks.
8. `sessions` - Numerical variable that indicates the total number of sessions of the treatment.
9. `minutes` - Numerical variable that indicates average time in minutes of each session of the treatment.
10. `math_area` - Character string that indicates the math area target of the treatment. Algebra (`algebra`), geometry (`geometry`), number (`number`), or multiple areas (`multiple_areas`).
11. `device` - Character string that indicates the device used during the treatment. Personal Computer (`pc`), console (`con`), or application on other device (`app`).
12. `n_effect` - Numerical variable that identifies within each study the different effects. In each study the first effect is identified with `1`, then the value increases if more effects are reported within the same study.
13. `id_effect` - Numerical variable that identifies the unique effect. Values range from `1` to `43`.
14. `dependence` - Character string that indicates the type of dependence among different effect sizes reported within the same study. Study with single effect size (`none`), multiple effects measured on the same participants (`multiple_outcomes`), multiple comparison between different experimental groups and same control group (`multiple_groups`), or independent effect sizes (`independent`).

[5]In the case of unpublished dissertations or grey-literature the year refers to the realization of the study

15. `N` - Numerical variable that indicates the total number of participants included to compute the considered effect. This is not always equal to the total number of participants in the study, as multiple experimental or control groups cold be included in the study.
16. `n_cg` - Numerical variable that indicates the number of participants in the control group for the considered effect.
17. `n_eg` - Numerical variable that indicates the number of participants in the experimental group for the considered effect.
18. `mot` - Numerical variable that indicates if the considered effect measured the motivation in terms of expectancy (`1`), in terms of value (`2`), or did not differentiate between the two aspects (`NA`).
19. `m_t1_cg` - Numerical variable that indicates the control group mean score in the pre-test.
20. `sd_t1_cg` - Numerical variable that indicates the control group standard deviation in the pre-test.
21. `m_t2_cg` - Numerical variable that indicates the control group mean score in the post-test.
22. `sd_t2_cg` - Numerical variable that indicates the control group standard deviation in the post-test.
23. `m_t1_eg` - Numerical variable that indicates the experimental group mean score in the pre-test.
24. `sd_t1_eg` - Numerical variable that indicates the experimental group standard deviation in the pre-test.
25. `m_t2_eg` - Numerical variable that indicates the experimental group mean score in the post-test.
26. `sd_t2_eg` - Numerical variable that indicates the experimental group standard deviation in the post-test.

The dataset is loaded in R and its structure is presented below.

```
data_raw <- read.csv(file_path, sep = ";", header = T, stringsAsFactors = F)
```

```
str(data_raw)
```

```
## 'data.frame': 43 obs. of  25 variables:
##  $ study     : int  1 2 2 2 3 3 3 4 5 6 ...
##  $ id        : int  54 961 961 961 432 432 432 436 982 983 ...
##  $ author    : chr  "Bai et al." "Ke" "Ke" "Ke" ...
##  $ year      : int  2012 2006 2006 2006 2008 2008 2008 2010 2018 2015 ...
##  $ pub       : chr  "yes" "no" "no" "no" ...
##  $ grade     : int  2 1 1 1 1 1 1 2 1 1 ...
##  $ weeks     : int  18 4 4 4 4 4 4 18 7 10 ...
##  $ sessions  : int  NA 8 8 8 8 8 8 18 7 NA ...
##  $ minutes   : int  NA 45 45 45 40 40 40 30 45 30 ...
##  $ device    : chr  "pc" "pc" "pc" "pc" ...
##  $ n_effect  : int  1 1 2 3 1 2 3 1 1 1 ...
##  $ id_effect : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ dependence: chr  "none" "independent" "independent" "independent" ...
```

```
##  $ N          : int  437 115 114 129 76 77 79 193 80 1168 ...
##  $ n_cg       : int  192 58 63 60 36 36 36 76 40 526 ...
##  $ n_eg       : int  245 57 51 69 40 41 43 117 40 642 ...
##  $ mot        : int  NA NA NA NA NA NA NA NA 1 2 ...
##  $ m_t1_cg    : num  75.6 156.2 149.4 145 77.5 ...
##  $ sd_t1_cg   : num  11.97 23.84 30.12 26.91 9.38 ...
##  $ m_t2_cg    : num  72.4 156.1 145.7 144.9 74.2 ...
##  $ sd_t2_cg   : num  11 26.5 28.9 28 12 ...
##  $ m_t1_eg    : num  71.2 150.4 142.1 152 77.7 ...
##  $ sd_t1_eg   : num  12.3 30.5 33.1 24.9 13.2 ...
##  $ m_t2_eg    : num  71.8 151.4 145.8 162 77.3 ...
##  $ sd_t2_eg   : num  13.5 29.7 30.2 23.9 15.6 ...
```

3.2 Data munging

Following `drake` guidelines, we define function for each step of the analysis. To organize the dataset we define the following function that allows to state which variables are categorical variables (factors) and uses more explicit labels in some variables. Moreover, we create a variable `author_y` to obtain a label in the format *"Author's (year)"* for each study and the variable `intensity = sessions*minutes/weeks` as a rough measure of the weekly intensity of each treatment.

```
clean_data <- function(data){
  data%>%
    # Define which variable are factor
    mutate_at(vars("study","id","pub","dependence","grade",
                   "math_area","device","mot"), factor)%>%
    # Redefine factor labels for grade and device
    mutate(grade=recode_factor(grade,"1"="Primary","2"="Secondary"),
           math_area=recode_factor(math_area,"multiple areas"="multiple_areas"),
           math_area=fct_relevel(math_area,"algebra","geometry","number","multiple_areas"),
           device=recode_factor(device,"app"="App","con"="Console","pc"="PC"),
           mot=recode_factor(mot, "1"="expectancy","2"="value"),
           # Obtain authors (year) label
           author_y=paste0(author," (",year,")"),
           # Compute intensity of the intervention
           intensity = sessions*minutes/weeks)%>%
    select("study":"minutes","intensity",everything())
}
```

In Pareto, Arvemo, Dahl, Haake, and Gulz (2011) the gain scores (i.e., the mean difference between post-test and pre-test scores) are reported instead of the post-test scores. Thus, to obtain the post-test scores we define a function that computes the post-test scores. The mean is given by the sum between the mean pre-test score and the mean gain score ($Mean_{post} = Mean_{pre} + Mean_{gain}$). The post-test Standard deviation is given by the square-root of the sum between the variance of the pre-test scores and the variance of the gain scores ($SD_{post} = \sqrt{SD^2_{pre} + SD^2_{gain}}$).

```r
pareto_post <- function(data){
  # Control group gain
  m_gain_cg<-c(1.5,-.23)     # Mean gain
  sd_gain_cg<-c(3.82,3.62)   # Sd gain

  # Experimental group gain
  m_gain_eg<-c(.93,1.44)     # Mean gain
  sd_gain_eg<-c(3.09,3.52)   # Sd gain

  data %>%
    mutate_cond(study=="11",  # Change only rows satisfying the condition
                # Control group
                m_t2_cg = m_t1_cg + m_gain_cg,
                sd_t2_cg = sqrt(sd_t1_cg^2+sd_gain_cg^2),
                #Experimental group
                m_t2_eg = m_t1_eg + m_gain_eg,
                sd_t2_eg = sqrt(sd_t1_eg^2+sd_gain_eg^2))
  }
```

3.3 Compute d_{ppc2} value and variance

To compute the d_{ppc2} value and variance, we use the functions reported in the `metafor` website (http://www.metafor-project.org/doku.php/analyses:morris2008). To compute the variance we need the correlation between pre-test and post-test scores. However, correlation are not reported so we try with different values high correlation (`r_high = .8`), medium-high correlation (`r_mediumh = .6`), medium-low correlation (`r_mediuml = .4`), and low correlation (`r_low = .2`).

```r
compute_dppc2 <- function(data){

  data%>%
    mutate(
      # compute pooled standard deviation
      sd_pool= sqrt(((n_eg-1)*sd_t1_eg^2+(n_cg-1)*sd_t1_cg^2)/(n_eg+n_cg-2)),
      # Compute yi_dppc2
      yi_dppc2 = metafor:::.cmicalc(n_eg+n_cg-2)*((m_t2_eg - m_t1_eg) - (m_t2_cg-m_t1_cg))
                                    /sd_pool,
      # add correlations pre-post
      r_high=.8,      # high correlation
      r_mediumh=.6,   # medium-high correlation
      r_mediuml=.4,   # medium-low correlation
      r_low=.2)%>%    # low correlation

    # Rearrange dataset from wide format to long format
    gather("r_high":"r_low", key="r_size",value="r_value", factor_key = TRUE)%>%
    # compute variance
    mutate(vi_dppc2=2*(1-r_value)*(1/n_eg + 1/n_cg)+ yi_dppc2^2/(2*(n_eg + n_cg)))%>%
    arrange(study,n_effect,r_size)  # ordering dataset
}
```

Simply looking at the effect obtained, we observe a strange value. In Hung, Huang, and Hwang (2014) values of `yi_dppc2` for the second effect are extremely huge 38.2. This value is impossible.

```
##   study  id      author year n_effect id_effect        dependence  N n_cg n_eg
## 1    13 959 Hung et al. 2014        1        32 multiple_outcomes 46   23   23
## 2    13 959 Hung et al. 2014        2        33 multiple_outcomes 46   23   23
##   m_t1_cg sd_t1_cg m_t2_cg sd_t2_cg m_t1_eg sd_t1_eg m_t2_eg sd_t2_eg
## 1     3.3     0.76    3.24     0.82    3.53     0.62    3.74     0.79
## 2     3.0     0.02    3.23     0.72    3.00     0.00    3.78     0.72
##      sd_pool   yi_dppc2    r_size r_value    vi_dppc2
## 1 0.69354164  0.3826259 r_mediumh     0.6  0.07115655
## 2 0.01414214 38.2235343 r_mediumh     0.6 15.95041926
```

Considering the measure in the pre-test we can understand the problem. The standard deviation in the pre-test are extremely low: 0.02 and 0. This leads to a low value for the pooled standard deviation and in turns an implausible high value of effect size. That is probably given by the fact that the measure used was not able to evaluate properly the variation among individuals in the pre-test, (maybe floor effect).

We do not include this effect in the meta-analysis and we remove it from the dataset. Thus the final dataset includes 42 different effects.

```
rm_hung <- function(data) {
    data %>% filter(study != "13" | n_effect != 2)
}
```

4 Descriptive Statistics

In this section we describe the main characteristics of the studies included in the meta-analysis. First we present descriptive statistics considering the study level, then we present descriptive statistics considering the effect sizes level.

4.1 Studies characteristics

In the analysis we include 19 studies and 42 different effects. Studies were published between 2006 and 2018 and most of the studies are published after 2010. The number of studies according to year of publication is presented in Figure 1.

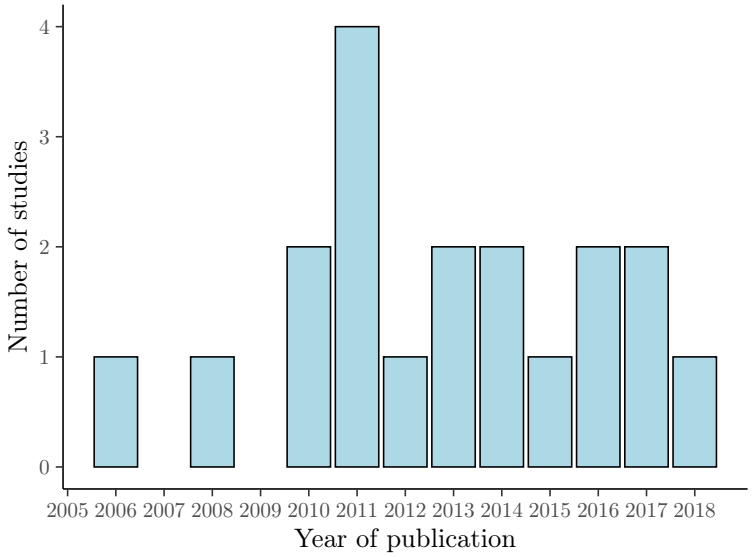

Figure 1: Number of studies according to year of pubblication ($n_{studies} = 19$).

The majority of the studies were published in a peer reviewed journal, but 6 studies were obtained from conference papers or unpublished dissertations (see Table 1).

Table 1: Type of publication.

Publication	Frequency
Peer-reviewed journal	13
Conference papers or dissertations	6

Note: $n_{studies} = 19$

The total number of participants included in each study range from 40 to 1168 and the majority of the studies included less than 150 participants. The number of participants for each study is presented in Figure 2

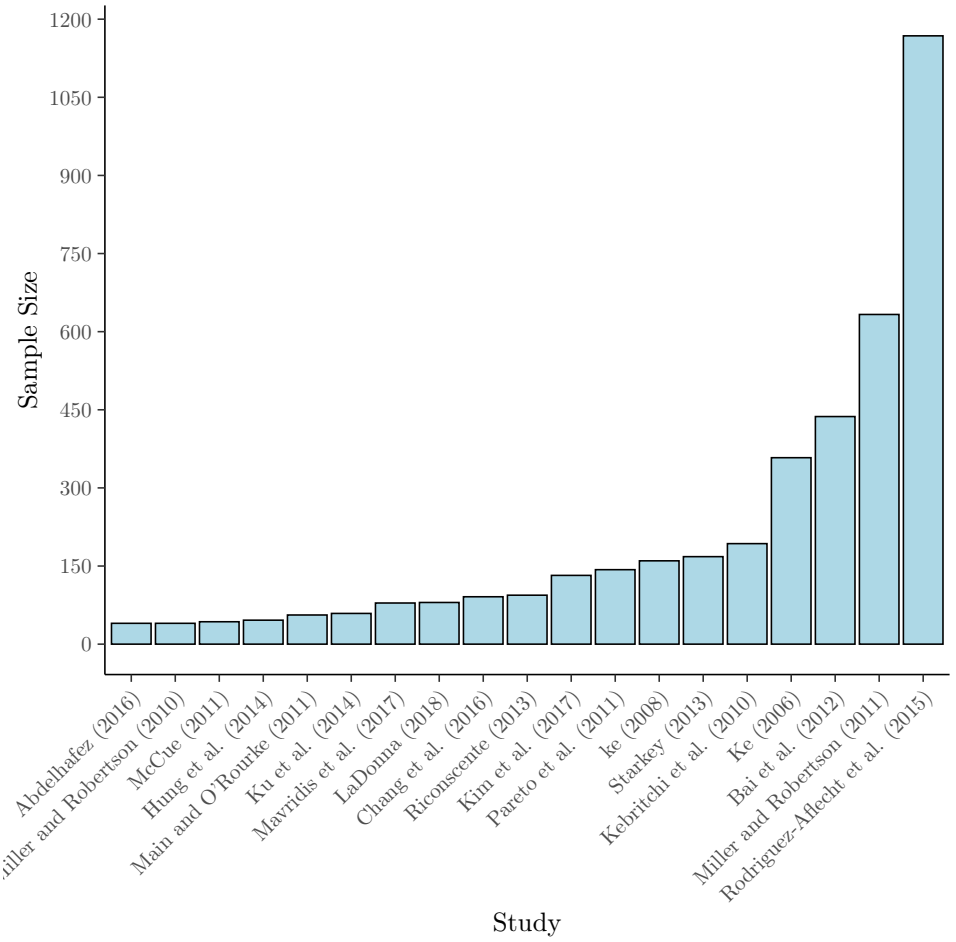

Figure 2: Number of participants included in each study ($n_{studies} = 19$).

The studies included in the meta-analysis differ according to the school grade of the participants. In more than a half of the studies, participants were primary school students and in the other studies participants were secondary school students (see Table 2).

Table 2: Participants school grade.

School grade	**Frequency**
Primary school	11
Secondary school	8

Note: $n_{studies} = 19$

Regarding the math area target of the treatment, 9 studies focused on numbers, 4 studies on algebra, only one study on geometry and integer(0) focused on multiple areas (see Table 3).

Table 3: Math area target of the treatment.

Math area	Frequency
Algebra	4
Geometry	1
Number	9
Multiple areas	5

Note: $n_{studies} = 19$

Considering the device used in the treatment, the large part of the studies used the Personal Computer. Only 3 studies used different type of console and 3 studies involved the use of Applications (see Table 4).

Table 4: Device used in the treatment.

Device	Frequency
App	3
Console	3
PC	13

Note: $n_{studies} = 19$

The duration of the treatment range from 1 week to 28 weeks. In Table 5 studies are grouped according to approximate duration of the treatment.

Table 5: Duration of treatment in weeks.

Duration	Frequency
1 week	3
4 weeks or less	3
10 weeks or less	9
more than 10 weeks	4

Note: Maximum value is 28 weeks; $n_{studies} = 19$

In Table 6 the frequency of studies according to the number of effects is reported. Th majority of the studies include only one effect, all the other studies include between 2 and 4 effects, except one study with 8 effects.

Table 6: Frequency of studies according to the number of effects.

Effects within study	Number of studies
1	10
2	3
3	2
4	3
8	1

Note: $n_{studies} = 19$; $n_{effects} = 42$

Studies with multiple effect size differ according to the type of dependency between effects. In 8 studies outcomes where measured on the same participants. Only 1 study included multiple independent treatment groups compared to a single control group and another study included multiple independent treatment groups and independent control groups (see Table 7).

Table 7: Type of dependency among effect sizes within the same study.

Dependency	Frequency
Independent experimental and control groups	1
Independent experimental groups same control	1
Multiple outcomes on same participants	8
Single effect	9

Note: $n_{studies} = 19$

4.2 Effect characteristics

In Figure 3 the estimated effect sizes of each study are presented together with confidence intervals. The studies are ordered according to the total number of participants included and multiple outcomes from the same study are grouped together. We can observe the high variability between effects and how with the increasing of the sample size effects tend to be smaller.

Considering the different measures of motivation, 13 effect sizes considered motivation in terms of expectancy and 19 effect sizes considered motivation in terms of value. The remaining effect sizes did not differentiate between the two aspects (see Table 8).

Table 8: Effect sizes according to the type of motivation measure.

Motivation measure	Frequency
Expectancy	13
Value	19
Not differentiated	10

Note: $n_{effects} = 42$

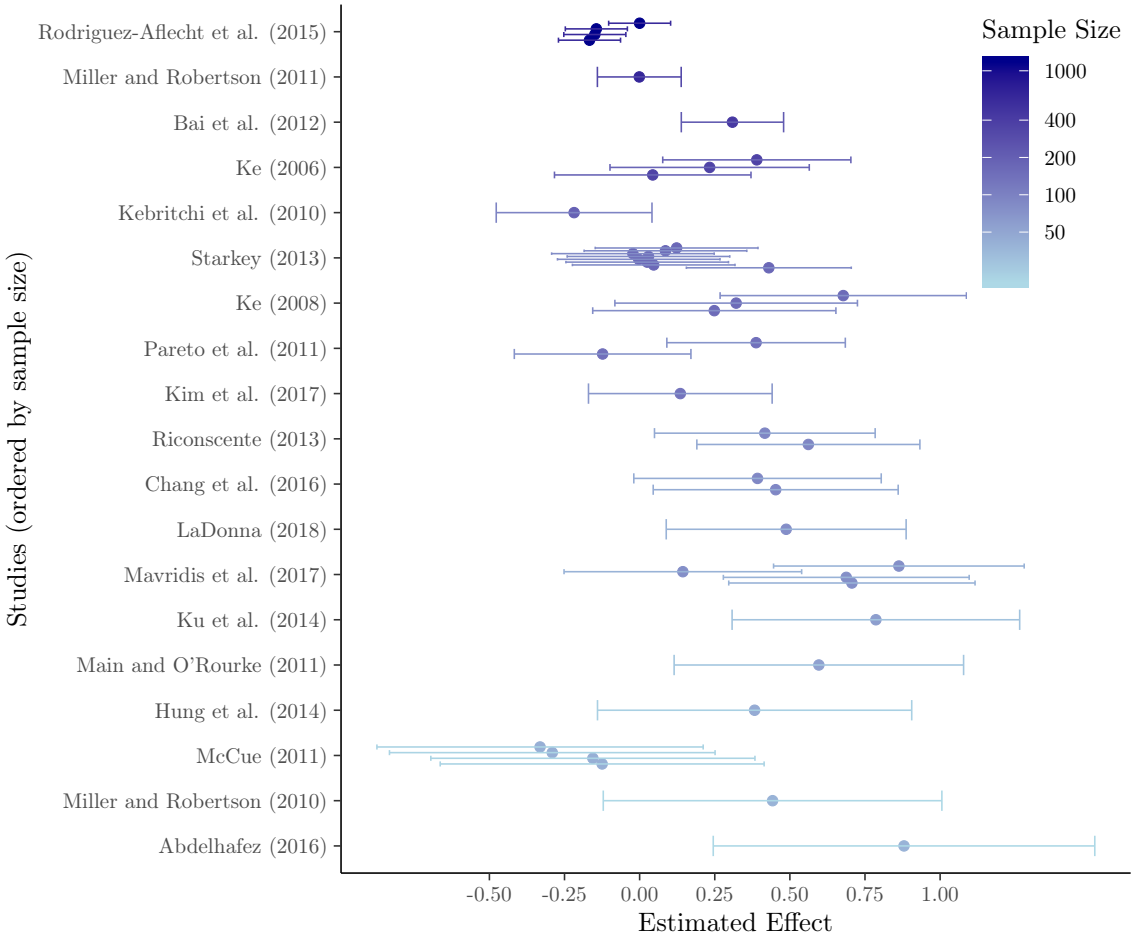

Figure 3: Estimated effect size according to number of participants ($n_{studies} = 19$; $n_{effects} = 42$).

5 Meta-Analysis Results

In this section, first we present the results of the meta-analysis. Then sensitivity analyses are performed to evaluate consistency of the results and the possible presence of publication bias. Finally, moderators analysis is described.

5.1 Main results meta-analysis

5.1.1 Variance-covariance matrix

In order to compute a multilevel meta-analysis we need first to define for each study the variance-covariance matrix according to the specific dependence structure between effect sizes.

As we have seen in Table 7 only 10 studies include multiple effect size. In the case of two outcomes (A and B) measured on the same subjects, the covariance between outcomes ($cov(AB)$) can be computed knowing the variance of the two outcomes (σ_A^2 and σ_B^2) and their correlation (ρ_{AB}):

$$cov(AB) = \rho_{AB} * \sigma_A \sigma_B. \tag{7}$$

The same formula applies for each couple of outcomes when multiple outcomes were measured. The variance-covariance matrix (Σ) can be computed as the product between the diagonal matrix that includes the outcomes standard deviation ($diag(S)$) and the correlation matrix ($Corr$): (Todo: Control notation)

$$\Sigma = diag(S) * Corr * diag(S). \tag{8}$$

However, none of the studies reported the correlation between values. Thus, variance-covariance matrix is computed with a range of different correlation values between outcomes and influence on the results is evaluated.

To compute the variance-covariance matrix in all studies, we used the `impute_covariance_matrix()` function included in the R-package `clubSandwich` (Pustejovsky, 2019). Given an assumed correlation value, the vector with all variances, and the grouping variable (in this case `study` as effects are grouped within studies), this function returns a list with the variance-covariance matrix for each study. 8 studies outcomes where measured on the same participants. In the case of studies with only one effect, the `impute_covariance_matrix()` function returns a 1x1 matrix with sampling variance value.

Ke (2008) is the only study where multiple independent treatment groups were compared to a common control group. In this case, covariance has to be computed considering this type of dependency between effects. Gleser and Olkin (2009) provided the formula to correctly compute the variance and covariance values for Cohen's d effect size when multiple experimental groups are compared to a common control group[6]. However, to our knowledge, no formula is available in this case for the d_ppc2 proposed by Morris (2008). Thus, we treat this study in the same way as studies with multiple measures on the same subjects. We allow correlation to vary in order to evaluate its influence on the results.

[6]The implemented R-functions are documented on the `metafor` website (http://www.metafor-project.org/doku.php/analyses:gleser2009)

Finally, Ke (2006) is the only study where multiple independent treatment groups were compared to multiple independent control groups. In this case the correlation between outcomes is set to zero.

We specify a function that given a correlation value computes and adjust the variance-covariance matrix for all studies according to dependency between effects.

```
compute_vcv_matrix <- function(data, r = 0.5) {

    # Considering all effects correlated
    cov_dppc2 = with(data, clubSandwich::impute_covariance_matrix(vi = vi_dppc2,
        cluster = study, r = r))

    # Effect sizes in Ke (2006) are independent so we set covariances to 0
    cov_dppc2[["2"]] = cov_dppc2[["2"]] * diag(1, 3, 3)  # Study id is '2'

    return(cov_dppc2)
}
```

5.1.2 Multilevel meta-analysis

Based on the `rma.mv()` function from `metafor` R-package, we define a function to fit the multilevel meta-analysis. First we filter the dataset considering a give pre- post-test correlation (`r_pre_post` can be set to `"r_high"`,`"r_mediumh"`, `"r_mediuml"`, `"r_low"`) used to compute the effect sampling variance (`vi_dppc2`, see Section 3.3). The variance-covariance matrix is computed according to a given correlation value between outcomes (`r_outcomes`). Then, the Multilevel meta-analysis is fitted considering the study clusters as a random-effects (`random = ~ 1|study`).

In the outcome we add the I^2 value computed following the functions presented on the `metafor` website (http://www.metafor-project.org/doku.php/tips:i2_multilevel_multivariate; see Section *"Multilevel Models"*) and the coefficient test using the `coef_test` from `clubSandwich` R-package that applies small sample correction. The function includes also two extra arguments that allow to exclude single studies and define moderators. These settings will be used for LOO sensitivity analysis and moderator analysis.

```
rma_multilevel <- function(data, r_pre_post = "r_mediumh", r_outocomes = 0.5,
    excluded_study = NULL, moderator = NULL) {

    # Filter data
    data = data %>% filter(r_size == r_pre_post)

    # compute variance-covariance matrix
    cov_dppc2 = compute_vcv_matrix(data, r = r_outocomes)

    # check if study have to be excluded (for the sens_loo)
    if (!is.null(excluded_study)) {
        # remove study from variance-covariance matrix
        cov_dppc2 = compute_vcv_matrix(data, r = 0.5)[-excluded_study]
        # remove study from data
        data = data %>% filter(study != excluded_study)
```

```
    }

    # Get mod formula
    if (is.null(moderator)) {
        mod_formula = NULL
    } else {
        mod_formula = as.formula(paste0("~", moderator))
    }

    # multilevel meta-analysis
    fit_rma_mv = rma.mv(yi = yi_dppc2, V = cov_dppc2, random = ~1 | study,
        mod = mod_formula, method = "REML", data = data, slab = author_y)

    # add useful information
    fit_rma_mv$I_squared = I_squared(fit_rma_mv)
    fit_rma_mv$coef_test = coef_test(fit_rma_mv, cluster = data$study, vcov = "CR2")
    fit_rma_mv$r_pre_post = r_pre_post
    fit_rma_mv$r_outocomes = r_outocomes
    fit_rma_mv$data = data
    fit_rma_mv$excluded_study = excluded_study

    return(fit_rma_mv)
}
```

Default settings consider `r_pre_post = "r_mediumh"`, which means a correlation of 0.6, and `r_outcomes = 0.5`. These are reasonable values researchers would expect. The influence of different values on the results is considered later. Below we report the output of the model `fit_rma_mv` based on default settings.

```
summary(fit_rma_mv)
```

```
##
## Multivariate Meta-Analysis Model (k = 42; method: REML)
##
##   logLik  Deviance       AIC       BIC      AICc
## -10.8182   21.6364   25.6364   29.0636   25.9522
##
## Variance Components:
##
##             estim    sqrt  nlvls  fixed  factor
## sigma^2    0.0574  0.2396     19     no   study
##
## Test for Heterogeneity:
## Q(df = 41) = 150.2925, p-val < .0001
##
## Model Results:
##
## estimate      se    zval    pval   ci.lb   ci.ub
##   0.2532  0.0672  3.7658  0.0002  0.1214  0.3849  ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model I^2 is 79% and the coefficients test with small sample correction are reported in Table 9

Table 9: Coefficients test for `fit_rma_mv`

Coefficient	Estimate	SE	t-value	df	p-value	Sig.
Intercept	0.25	0.07	3.77	16.76	0.0016	**

Note: $n_{studies} = 19$; $n_{effects} = 42$

The results indicate that the treatments have a significant small-effect ($d_{ppc2} = 0.25$). However, the confidence interval ($95\% CI = [0.12; 0.38]$) range from very small-effects to medium effects indicating that there is a high level of uncertainty in the estimation. Moreover, results show high levels of heterogeneity in the studies included in the meta-analysis. The Q-statistic is significant $Q(\text{df} = 41) = 150.29$ and I^2 value is high ($I^2 = 79\%$). The Forest plot is presented in Figure 4.

5.1.3 Robust variance analysis

In order to evaluate the result we would obtain using the Robust Variance Analysis (RVA) approach, we fit a different model using the function `robu()` from the `robumeta` R-package. We consider a correlated effects dependence structure for the analysis (`modelweights = "CORR"`) and the same correlation between outcomes as in the previous analysis (`rho=.5`)

```
rva_meta <- function(data) {
    # Robust Variance Analysis
    robumeta::robu(yi_dppc2 ~ 1, data = data %>% filter(r_size == "r_mediumh"),
        modelweights = "CORR", studynum = study, var.eff.size = vi_dppc2,
        small = TRUE, rho = 0.5)
}
```

Results are presented below. Overall we have similar results to the previous analysis. The treatments effect is small and the heterogeneity between studies is high.

```
## RVE: Correlated Effects Model with Small-Sample Corrections
##
## Model: yi_dppc2 ~ 1
##
## Number of studies = 19
## Number of outcomes = 42 (min = 1 , mean = 2.21 , median = 1 , max = 8 )
## Rho = 0.5
## I.sq = 77.24818
## Tau.sq = 0.06526261
##
##                Estimate StdErr t-value  dfs P(|t|>) 95% CI.L 95% CI.U Sig
## 1 X.Intercept.    0.261 0.0692    3.77 16.9 0.00155    0.115    0.406 ***
## ---
## Signif. codes: < .01 *** < .05 ** < .10 *
## ---
## Note: If df < 4, do not trust the results
```

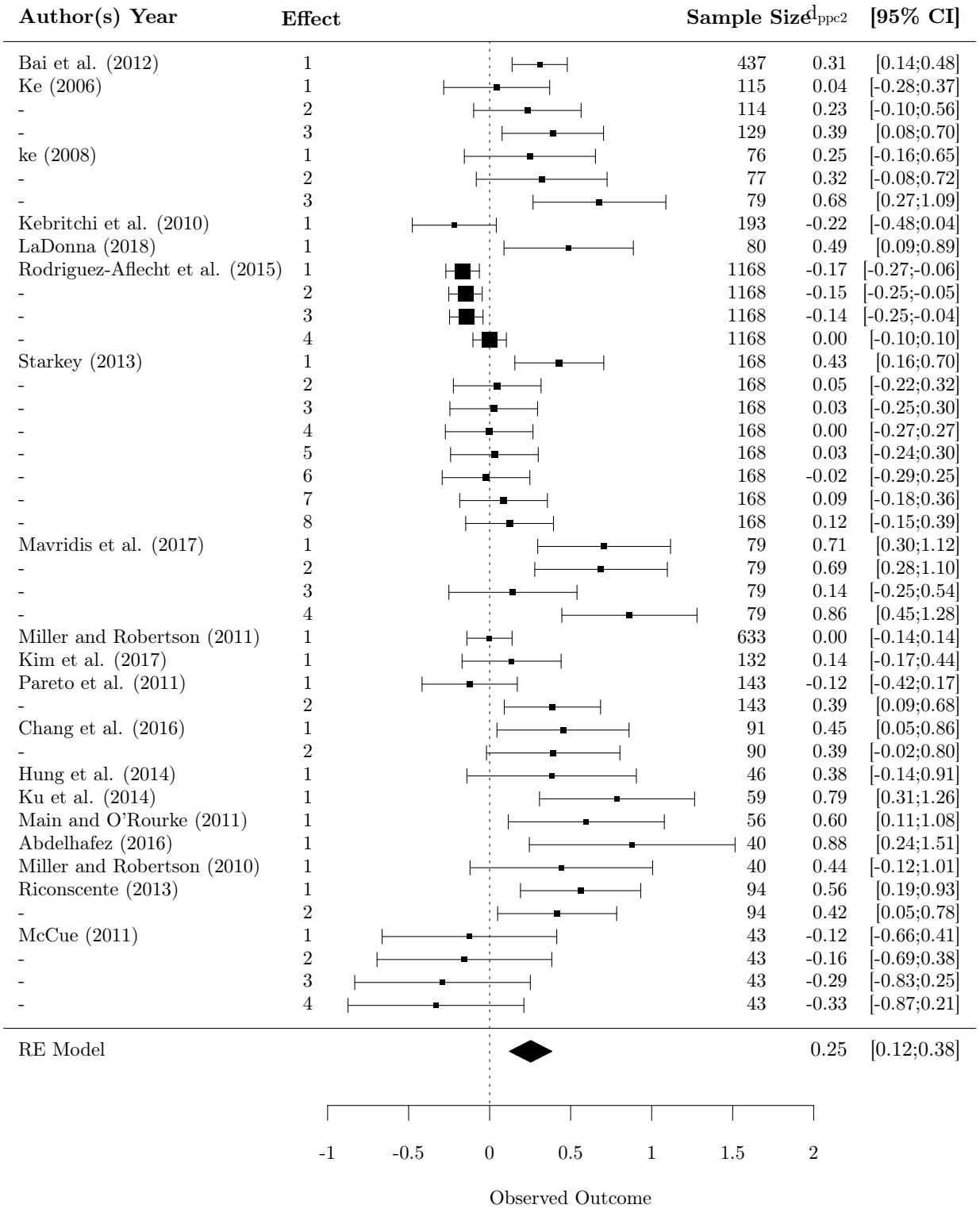

Figure 4: Forest plot of the multilevel meta-analysis ($n_{studies} = 19$; $n_{effects} = 42$).

5.1.4 Aggregated effects meta-analysis

In order to evaluate the result we would obtain averaging effects within studies, we conduct a separate meta-analysis. First, we compute the composite effect for each study using `agg()` function from `MAd` R-package (Del Re & Hoyt, 2014). Following, Hoyt and Del Re (2018) recommendations, we set `method="BHHR"` to use the approach suggested by Borenstein et al. (2009) and we set correlation between outcomes at 0.5 as in the previous cases.

```
aggregate_data = function(data, r_pre_post = "r_mediumh", cor = 0.5, method = "BHHR") {
    selected_data = data %>% filter(r_size == r_pre_post)

    # Aggregate effects
    agg_effects = MAd::agg(id = study, es = yi_dppc2, var = vi_dppc2, cor = cor,
        n.1 = n_eg, n.2 = n_cg, method = method, data = selected_data)

    # Arrenge dataset to keep only useful variables
    data_aggregated = selected_data %>% group_by(study) %>% mutate(N = round(mean(N),
        0), n_cg = round(mean(n_cg), 0), n_eg = round(mean(n_eg), 0)) %>%
        filter(!duplicated(study)) %>% left_join(., agg_effects, by = "id") %>%
        mutate(yi_dppc2 = es, vi_dppc2 = var) %>% select("study":"n_eg", "author_y",
        "yi_dppc2":"vi_dppc2")

    return(data_aggregated)
}
```

Then, we fit a random-effects model meta-analysis using `rma()` function from `metafor` R-package. Results are presented below. Overall we have similar results to the previous analysis. The treatments effect is small and the heterogeneity between studies is high.

```
##
## Random-Effects Model (k = 19; tau^2 estimator: REML)
##
##   logLik  deviance       AIC       BIC      AICc
##  -4.2653    8.5306   12.5306   14.3114   13.3306
##
## tau^2 (estimated amount of total heterogeneity): 0.0595 (SE = 0.0288)
## tau (square root of estimated tau^2 value):      0.2439
## I^2 (total heterogeneity / total variability):   79.62%
## H^2 (total variability / sampling variability):  4.91
##
## Test for Heterogeneity:
## Q(df = 18) = 88.1854, p-val < .0001
##
## Model Results:
##
## estimate      se    zval    pval   ci.lb   ci.ub
##   0.2569  0.0684  3.7572  0.0002  0.1229  0.3909  ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5.2 Sensitivity analyses

5.2.1 Sensitivity correlation values

To evaluate the influence of the correlation between pre-test and post-test scores (`r_pre_post`) and the influence of the correlation between outcomes (`r_outcomes`), we run the multilevel meta-analysis trying different combination of values. The d_{ppc2} sampling variance was previously computed considering high values (`r_pre_post`=0.8), medium-high values (`r_pre_post`=0.6), medium.low values (`r_pre_post`=0.4), and low values (`r_pre_post`=0.2) for the pre- post-test correlation (see Section 3.3).

For all conditions, several multilevel meta-analyses are run varying the correlation between outcomes used to compute the variance-covariance matrix. The sequence of values considered for the `r_outcomes` range from 0.1 to 0.9 with increments of 0.1. Thus, in the sensitivity analysis 36 different models are fitted.

The results of each different combination are saved in the R-object `sens_summary` that can be loaded using the `drake` function `loadd(sens_summary)`. The summary results are reported in Table 10 and presented in Figure 5 (in the plot 95% CI are not represented as they are mostly overlapping and this would make the plot messy and difficult to read).

Table 10: Summary report of the correlations sensitivity analysis.

Parameter	Min	Mean	Max
σ			
Estimate	0.18	0.23	0.28
95% CI lb	0.08	0.13	0.19
95% CI ub	0.35	0.38	0.42
d_{ppc2}			
Estimate	0.21	0.24	0.28
95% CI lb	0.07	0.11	0.14
95% CI ub	0.34	0.37	0.41
Heterogeneity			
I^2	52.37	73.07	90.82

Overall, results of the meta-analysis are only slightly influenced by the choice of the `r_pre_post` and `r_outcomes` values. The standard deviation of the random effects of studies (σ) range from 0.18 to 0.28 with a mean value of 0.23. The global effect of the treatments (d_{ppc2}) ranges from 0.21 to 0.28 with a mean value of 0.24.

Grater values of `r_pre_post` (and therefore smaller values of effect sampling variance `vi_dppc2`, see Section 3.3) lead to greater values of σ, d_{ppc2}, and I^2. On the contrary, greater values of `r_outcomes` (and therefore greater values of variance in the estimations as the independent information in data is lower) lead to smaller values of σ, d_{ppc2}, and I^2.

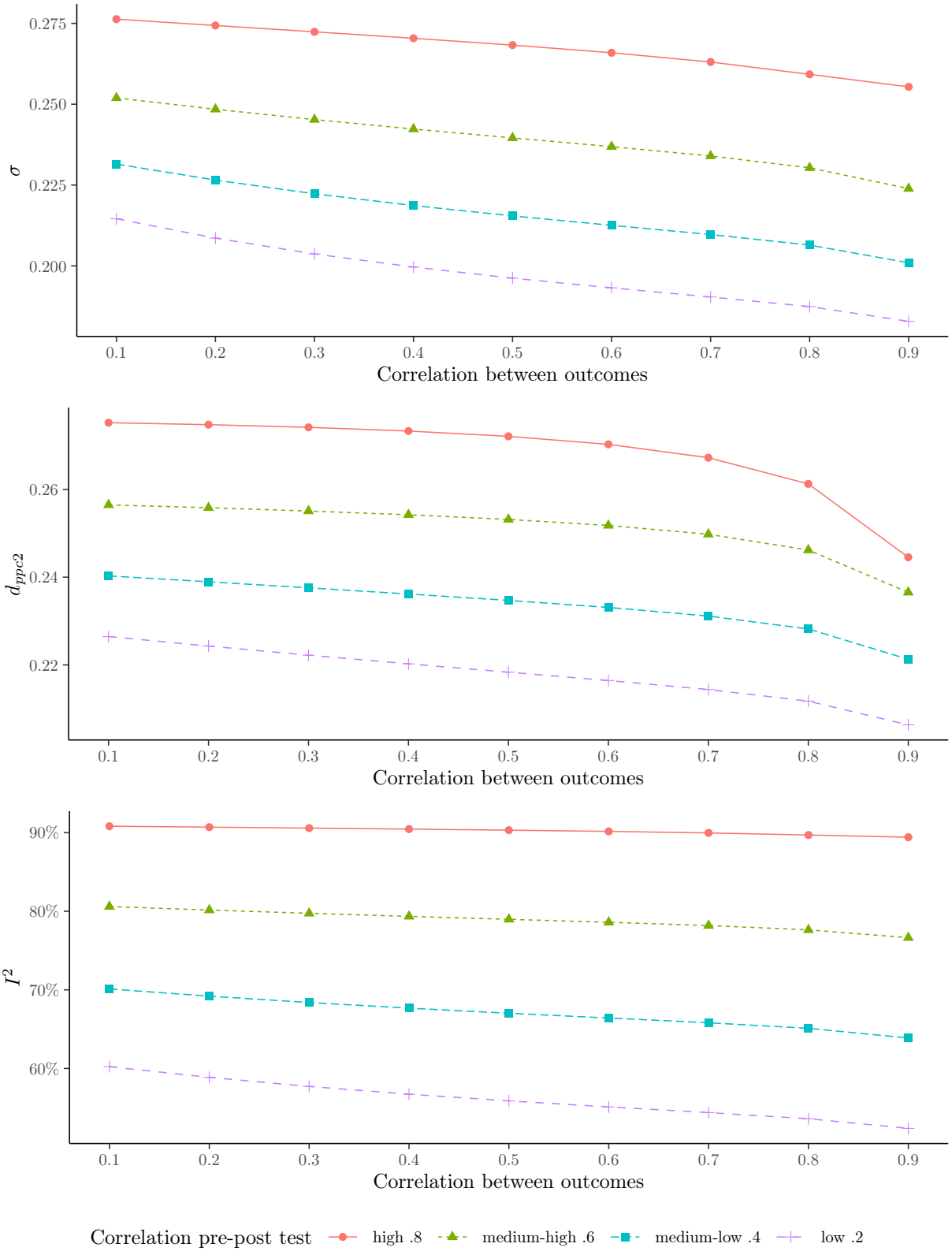

Figure 5: Results of the sensitivity analysis regarding `r_pre_post` and `r_outcomes`.

5.2.2 *Leave-One-Out* analysis

To assess the possible presence of influential studies, we re-run the multilevel meta-analysis removing one study at time, thus we fit 19 different models. The complete results of each different model are saved in the R-object `sens_loo_summary` that can be loaded using the `drake` function `loadd(sens_loo_summary)`. The summary results are reported in Table 11 and presented in Figure 6.

Table 11: Summary report of the LOO sensitivity analysis.

Parameter	Min	Mean	Max
σ			
Estimate	0.22	0.24	0.25
95% CI lb	0.12	0.14	0.15
95% CI ub	0.37	0.40	0.42
$\mathbf{d_{ppc2}}$			
Estimate	0.23	0.25	0.28
95% CI lb	0.10	0.12	0.15
95% CI ub	0.36	0.39	0.41
Heterogeneity			
I^2	64.39	78.61	81.22

Overall, results of the meta-analysis do not change considerably when one study is removed. The standard deviation of the random effects of studies (σ) range from 0.22 to 0.25 with a mean value of 0.24. The global effect of the treatments (d_{ppc2}) ranges from 0.23 to 0.28 with a mean value of 0.25.

Moreover, to evaluate the presence of influential studies we compute Cook's distance for each study (Cook, 1977). Cook's distances are presented in Figure 7. Rodríguez-Aflecht et al. (2015) has the highest Cook's distance ($D = 0.15$). To evaluate if a case is influential, several cut-offs have been proposed for Cook's distances. Few of these cut-offs are for example: $D_i > 1$; $D_i > 4/n$ where n is the number of cases; $D_i > 4/(n-k-1)$ where k is the number of model parameters.

However, as suggested by Fox (1991), a better option is to consider the graphical representation of Cook's distances to examine if there are cases that present a substantially larger distance than the rest. In Figure 7 we can observe a homogeneous distribution of distances. Thus, we can conclude that there are no influential cases.

5.3 Publication bias

To evaluate the presence of publication bias in the studies selected in the meta-analysis, the Funnel plot in presented in Figure 8. We can observe that the distribution of effects is not symmetric with respect to the estimated overall treatments effect (d_{ppc2}). Studies with larger sample size tend to report smaller effects, whereas studies with smaller sample size show greater effects.

However, this is not the optimal way to represent and evaluate the presence of publication bias, as studies with multiple effects are overrepresented in the plot respect to studies with only one effect.

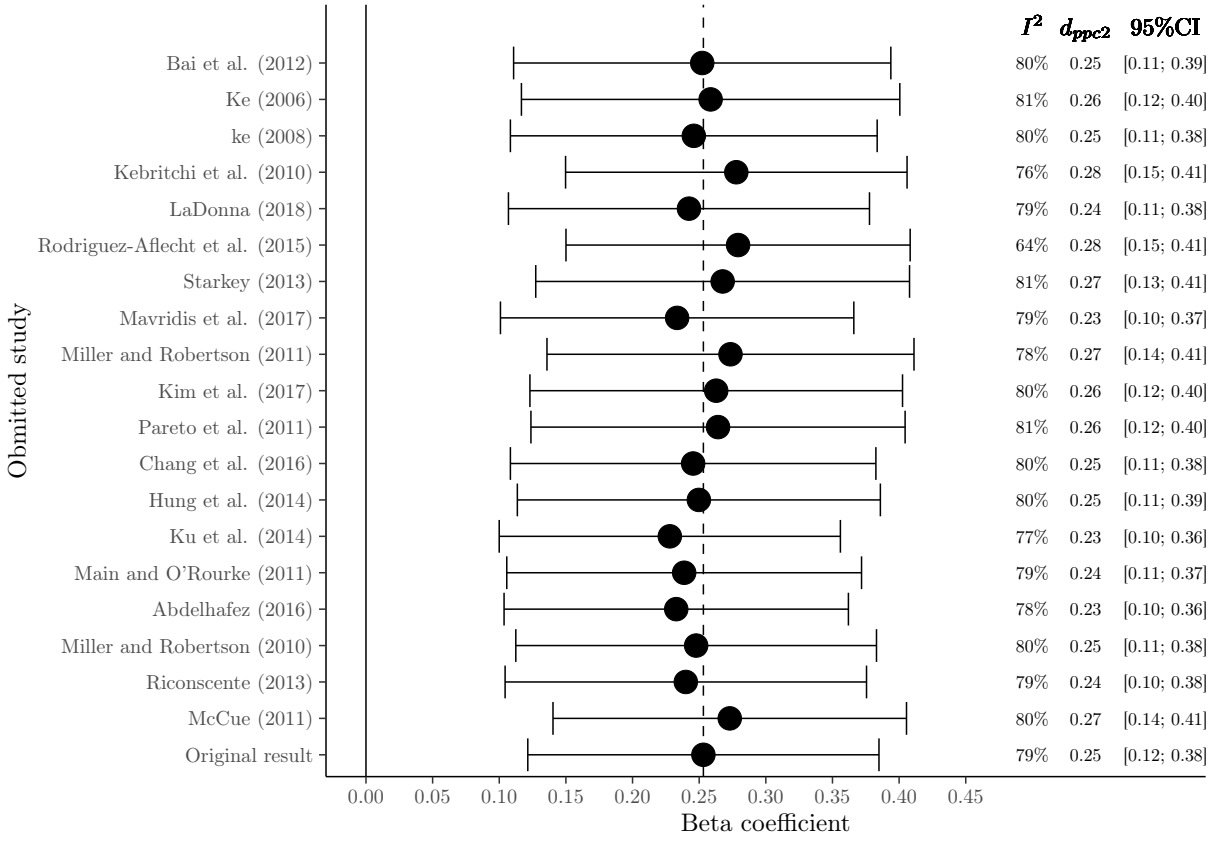

Figure 6: Results of the *leave-one-out* sensitivity analysis.

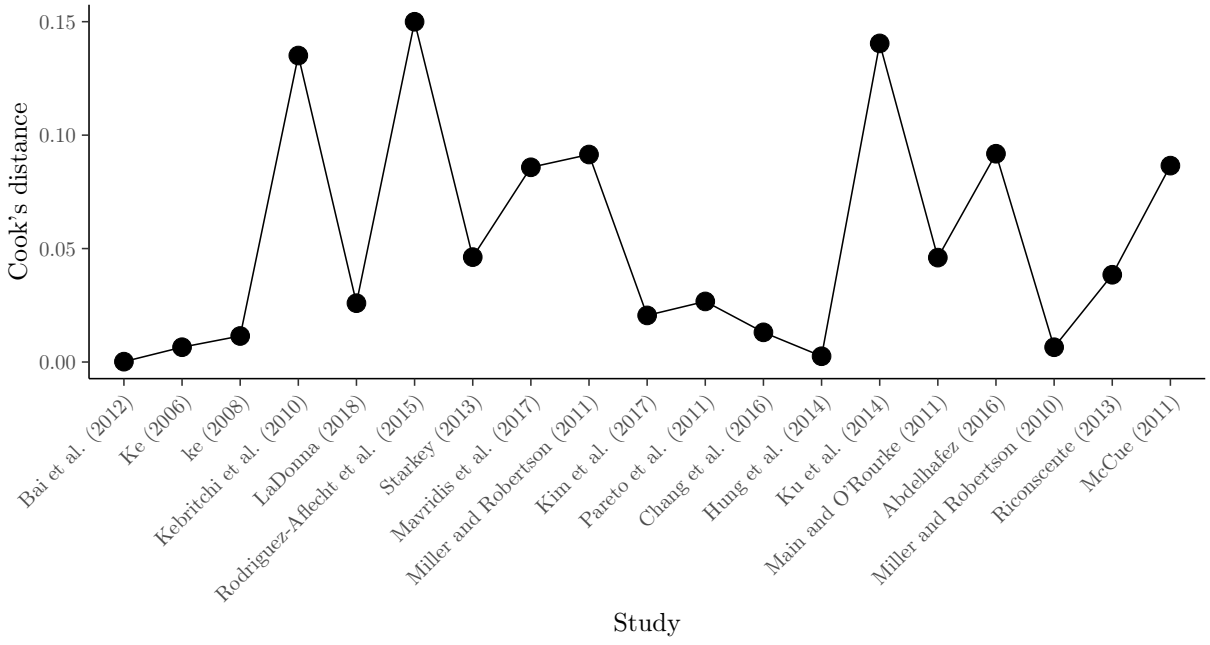

Figure 7: Cook's distance of each study ($n_{studies} = 19$).

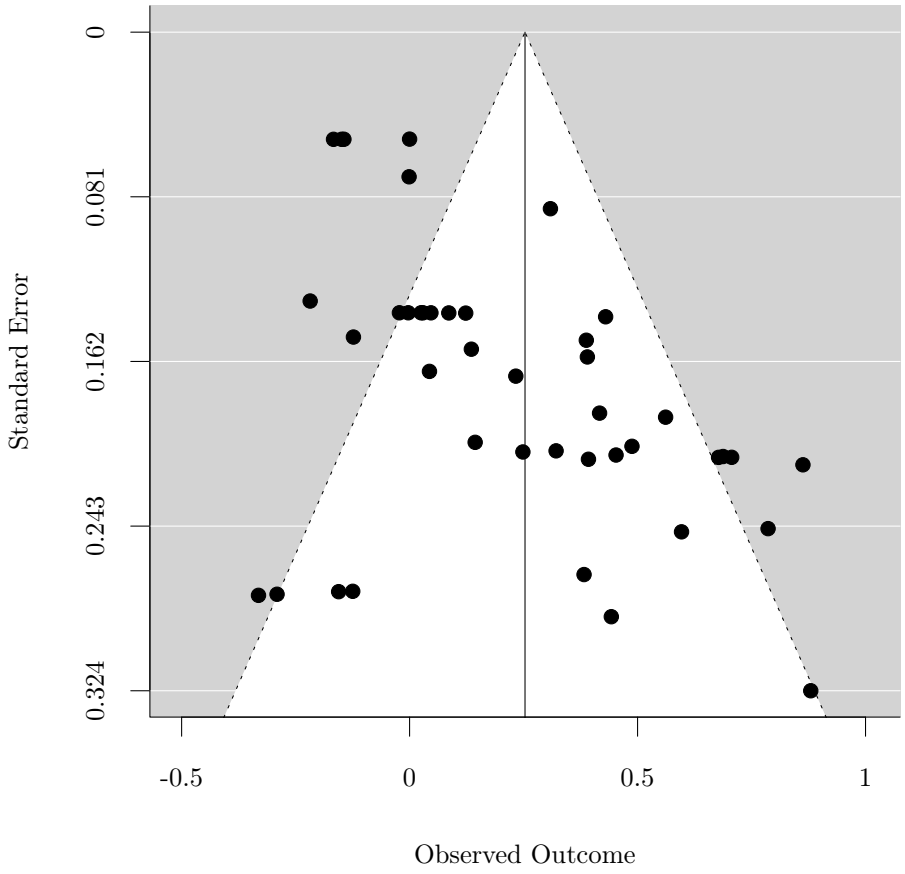

Figure 8: Funnel plot of the multilevel meta-analysis ($n_{studies} = 19$; $n_{effects} = 42$).

To overcome this issue we conduct two Egger's regression tests. The main idea is to evaluate if within studies the sample size (`N`) or the sampling variance (`vi_dppc2`) are associated to the estimated effect size (`yi_dppc2`). In absence of publication bias, studies results should be symmetrically distributed with respect to the estimated overall effect. Thus, studies results are expected to not be associated to sample size, sampling variance, or other indexes of the precision.

In the first Egger's regression test, we add the sample size as moderator in the multilevel meta-analysis (`mod = ~ N`). Results are presented below.

```
##
## Multivariate Meta-Analysis Model (k = 42; method: REML)
##
##   logLik  Deviance       AIC       BIC      AICc
##  -8.0941   16.1881   22.1881   27.2548   22.8548
##
## Variance Components:
##
##             estim    sqrt  nlvls  fixed  factor
## sigma^2    0.0356  0.1886     19     no   study
##
## Test for Residual Heterogeneity:
```

```
## QE(df = 40) = 105.2637, p-val < .0001
##
## Test of Moderators (coefficient 2):
## QM(df = 1) = 6.5060, p-val = 0.0108
##
## Model Results:
##
##          estimate      se     zval    pval    ci.lb    ci.ub
## intrcpt    0.3558  0.0736   4.8328  <.0001   0.2115   0.5001  ***
## N         -0.0004  0.0002  -2.5507  0.0108  -0.0008  -0.0001    *
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the second Egger's regression test, we add the sampling variance as moderator in the multilevel meta-analysis (`mod = ~ vi_dppc2`). Results are presented below.

```
##
## Multivariate Meta-Analysis Model (k = 42; method: REML)
##
##   logLik  Deviance       AIC       BIC      AICc
##  -6.2914   12.5828   18.5828   23.6494   19.2495
##
## Variance Components:
##
##             estim    sqrt  nlvls  fixed  factor
## sigma^2    0.0280  0.1674     19     no   study
##
## Test for Residual Heterogeneity:
## QE(df = 40) = 105.5386, p-val < .0001
##
## Test of Moderators (coefficient 2):
## QM(df = 1) = 11.2884, p-val = 0.0008
##
## Model Results:
##
##           estimate      se     zval    pval    ci.lb    ci.ub
## intrcpt    -0.0009  0.0865  -0.0101  0.9919  -0.1705   0.1687
## vi_dppc2    7.3137  2.1768   3.3598  0.0008   3.0472  11.5802  ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] -0.0004478989
```

In both cases the test of moderators is significant. For the sample size, we have $QM(df = 1) = 6.51$, $p - value = 0.011$ and the parameter coefficient ($\beta = -0.00045$) indicates that the expected effect size decreases as the study sample size increases. Note that even if the parameter value is small it refers to the effect of adding one participant, so for large sample size the changes may be remarkable.

For the sampling variance, we have $QM(df = 1) = 11.29$, $p - value = 0.001$ and the parameter coefficient ($\beta = 7.31$) indicates that, when the sapling variance decreases (i.e., more precise studies), the

expected effect size decreases as well.

In addition to the previous tests, we also report results of the rank correlation test that confirm the correlation between the observed outcomes and the corresponding sampling variances (Begg & Mazumdar, 1994).

```
ranktest(fit_rma_mv)

##
## Rank Correlation Test for Funnel Plot Asymmetry
##
## Kendall's tau = 0.3310, p = 0.0018
```

Finally, we present the funnel plot with the *trim-and-fill* method considering the results of the meta-analysis using aggregated effect sizes (see Section 5.1.4). The results are reported below and the and the funnel plot is represented in Figure 9.

```
##
## Estimated number of missing studies on the left side: 5 (SE = 2.9070)
##
## Random-Effects Model (k = 24; tau^2 estimator: REML)
##
## tau^2 (estimated amount of total heterogeneity): 0.0933 (SE = 0.0371)
## tau (square root of estimated tau^2 value):      0.3054
## I^2 (total heterogeneity / total variability):   84.24%
## H^2 (total variability / sampling variability):  6.35
##
## Test for Heterogeneity:
## Q(df = 23) = 108.2615, p-val < .0001
##
## Model Results:
##
## estimate      se    zval    pval   ci.lb   ci.ub
##   0.1514  0.0730  2.0747  0.0380  0.0084  0.2945  *
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results indicate that, to obtain symmetry, five hypothetical studies have to be added. The resulting treatments effect is still significant, but the effect size is very small ($d_{ppc2} = 0.15$).

Overall, the results indicate the presence of publication bias.

5.4 Moderators analysis

In this last part of the analysis, we evaluate he role of possible moderators such as type of publication (`mod = ~ pub`), school grade of the participants (`mod = ~ grade`), duration of the treatment in weeks (`mod = ~ weeks`), intensity of the treatment (`mod = ~ intensity`; see Section 3.2), math area target of the treatment (`mod = ~ math_area`), device used in the treatment (`mod = ~ device`), and type of motivation measured (`mod = ~ mot`).

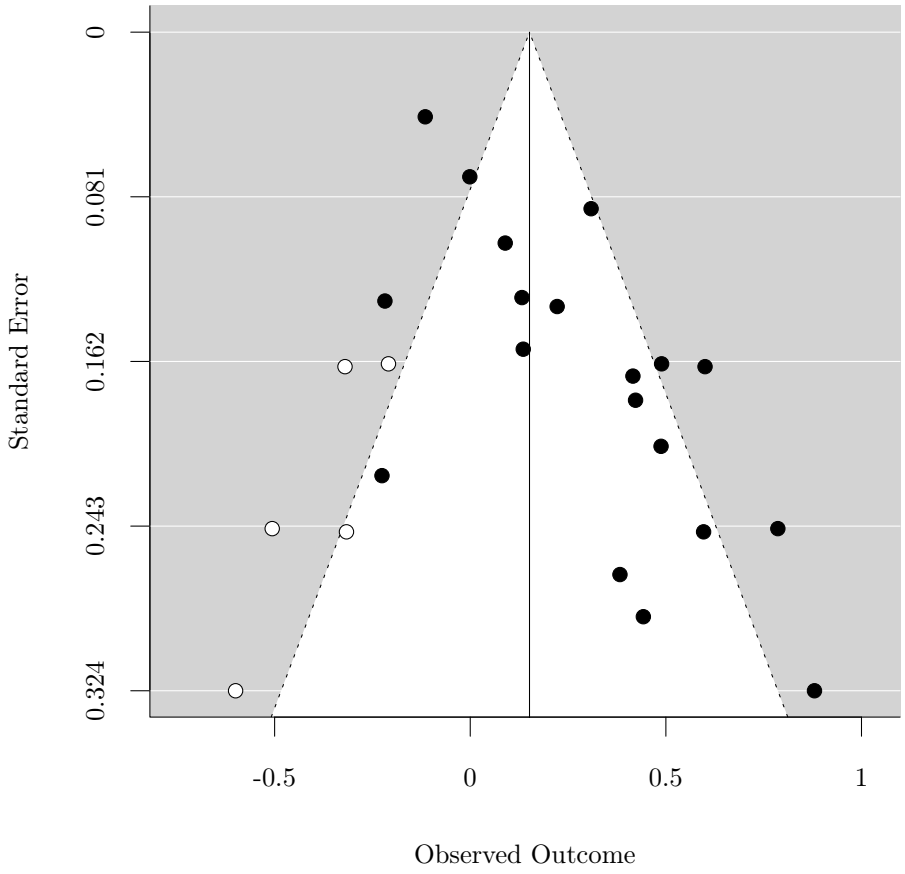

Figure 9: Funnel plot with *trim-and-fill* method for meta-analysis with aggregated effects (white dots are the added studies; $n_{studies} = 19$; $n_{effects} = 42$).

All the moderators are variables at the studies level. Only `mot` is a variable at the outcomes level as within the same study measures can differ according to the type of motivation. Three-level meta-analysis allows to investigate also this aspect that would not be possible to evaluate otherwise.

For each moderator a separate model is fitted Thus, results have to be interpreted with caution as moderators are not evaluated together to understand the unique variance explained by each moderator. Moreover, not all the studies include information about `intensity` and `mot`. In this cases the analysis is restricted to studies with complete data. Results of the moderator analysis are reported in Table 12.

Table 12: Results of the moderators analysis

Moderator	Studies	Effects	Q-value	df	p-value	Sign.
`pub`	19	42	0.17	1	0.678	
`grade`	19	42	1.06	1	0.303	
`weeks`	19	42	4.24	1	0.040	*
`intensity`	14	31	0.89	1	0.346	
`math_area`	19	42	5.78	3	0.123	
`device`	19	42	1.24	2	0.539	
`mot`	13	32	8.13	1	0.004	**

Looking at the test of moderators, we note that only the duration of the treatment in weeks and the type of motivation measured are significant moderators. To understand the direction of the effects we evaluate the model parameters.

Considering `weeks`, the effect of the treatment decreases as the number of weeks increases ($\beta = -0.02$). Complete results are reported below.

```
##
## Multivariate Meta-Analysis Model (k = 42; method: REML)
##
## Variance Components:
##
##             estim    sqrt  nlvls  fixed  factor
## sigma^2    0.0478  0.2186     19     no   study
##
## Test for Residual Heterogeneity:
## QE(df = 40) = 143.5199, p-val < .0001
##
## Test of Moderators (coefficient 2):
## QM(df = 1) = 4.2385, p-val = 0.0395
##
## Model Results:
##
##          estimate      se     zval    pval    ci.lb    ci.ub
## intrcpt    0.4257  0.1072   3.9732  <.0001   0.2157   0.6357  ***
## weeks     -0.0198  0.0096  -2.0588  0.0395  -0.0387  -0.0010    *
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##                [,1]
## intrcpt  0.42572938
## weeks   -0.01983365
```

Considering `mot`, the effect of the treatment is greater if motivation is measured in terms of expectancy than if motivation is measured in terms of value ($\beta = -0.1$). Complete results are reported below.

```
##
## Multivariate Meta-Analysis Model (k = 32; method: REML)
##
## Variance Components:
##
##             estim    sqrt  nlvls  fixed  factor
## sigma^2    0.0611  0.2473     13     no   study
##
## Test for Residual Heterogeneity:
## QE(df = 30) = 104.5503, p-val < .0001
##
## Test of Moderators (coefficient 2):
## QM(df = 1) = 8.1341, p-val = 0.0043
##
## Model Results:
##
```

```
##           estimate      se     zval    pval    ci.lb    ci.ub
## intrcpt     0.3078  0.0855   3.6019  0.0003   0.1403   0.4753  ***
## motvalue   -0.0981  0.0344  -2.8520  0.0043  -0.1655  -0.0307   **
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5.5 Conclusions

Overall, results indicate that there is a small effect of treatments. Moreover, treatments seems to have a larger effect on motivation in terms of `expectancy` than motivation in terms of `value` and the duration of the treatments seems to be a moderator factor, with shorter treatments having larger effects than longer ones.

However, this results have to be considered with caution given the great heterogeneity between studies and the presence of publication bias.

6 Session Information

```
sessionInfo(package = NULL)

## R version 3.6.1 (2019-07-05)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Mojave 10.14.6
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] grid      stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] tidyselect_0.2.5 robumeta_2.0     drake_7.8.0      metafor_2.1-0
##  [5] Matrix_1.2-17    kableExtra_1.1.0 forcats_0.4.0    stringr_1.4.0
##  [9] dplyr_0.8.3      purrr_0.3.3      readr_1.3.1      tidyr_1.0.0
## [13] tibble_2.1.3     ggplot2_3.2.1    tidyverse_1.2.1  knitr_1.26
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.3        lubridate_1.7.4   txtq_0.2.0        lattice_0.20-38
##  [5] png_0.1-7         utf8_1.1.4        assertthat_0.2.1  zeallot_0.1.0
##  [9] digest_0.6.23     R6_2.4.1          cellranger_1.1.0  backports_1.1.5
## [13] evaluate_0.14     httr_1.4.1        highr_0.8         pillar_1.4.3
## [17] rlang_0.4.2       lazyeval_0.2.2    readxl_1.3.1      rstudioapi_0.10
## [21] tikzDevice_0.12.3 rmarkdown_1.16    labeling_0.3      webshot_0.5.2
## [25] tinytex_0.16      igraph_1.2.4.2    munsell_0.5.0     broom_0.5.3
## [29] compiler_3.6.1    modelr_0.1.5      xfun_0.11         pkgconfig_2.0.3
## [33] htmltools_0.4.0   gridExtra_2.3     codetools_0.2-16  fansi_0.4.0
## [37] viridisLite_0.3.0 crayon_1.3.4      withr_2.1.2       nlme_3.1-141
## [41] jsonlite_1.6      gtable_0.3.0      lifecycle_0.1.0   magrittr_1.5
## [45] formatR_1.7       storr_1.2.1       scales_1.1.0      cli_2.0.0
## [49] stringi_1.4.3     farver_2.0.1      xml2_1.2.2        filelock_1.0.2
## [53] generics_0.0.2    vctrs_0.2.1       tools_3.6.1       glue_1.3.1
## [57] hms_0.5.2         colorspace_1.4-1  base64url_1.4     filehash_2.4-2
## [61] rvest_0.3.4       haven_2.1.1
```

Appendix A

Studies included in the meta-analysis

Abdelhafez, A. (2016). The effects of game-based technology on high school students' algebraic learning in an urban school classroom. *Available from Publicly Available Content Database.*

Bai, H., Pan, W., Hirumi, A., & Kebritchi, M. (2012). Assessing the effectiveness of a 3-D instructional game on improving mathematics achievement and motivation of middle school students. *British Journal of Educational Technology*, *43*(6), 993–1003. doi:10.1111/j.1467-8535.2011.01269.x

Chang, M., Evans, M. A., Kim, S., Norton, A., Deater-Deckard, K., & Samur, Y. (2016). The effects of an educational video game on mathematical engagement. *Education and Information Technologies*, *21*(5), 1283–1297. doi:10.1007/s10639-015-9382-8

Hung, C.-M., Huang, I., & Hwang, G.-J. (2014). Effects of digital game-based learning on students' self-efficacy, motivation, anxiety, and achievements in learning mathematics. *Journal of Computers in Education*, *1*(2-3), 151–166. doi:10.1007/s40692-014-0008-8

Ke, F. (2006). *Computer-based game playing within alternative classroom goal structures on fifth-graders' math learning outcomes: Cognitive, metacognitive, and affective evaluation and interpretation* (Doctoral dissertation, The Pennsylvania State University).

Ke, F. (2008). Alternative goal structures for computer game-based learning. *International Journal of Computer-Supported Collaborative Learning*, *3*(4), 429–445. doi:10.1007/s11412-008-9048-2

Kebritchi, M., Hirumi, A., & Bai, H. (2010). The effects of modern mathematics computer games on mathematics achievement and class motivation. *Computers & Education*, *55*(2), 427–443. doi:10.1016/j.compedu.2010.02.007

Kim, H., Ke, F., & Paek, I. (2017). Game-based learning in an OpenSim-supported virtual environment on perceived motivational quality of learning. *Technology, Pedagogy and Education*, *26*(5), 617–631. doi:10.1080/1475939X.2017.1308267

Ku, O., Wu, D., Lao, A., Wang, J., & Chan, T. (2014). The effects of mini-games on students' confidence and performance in mental calculation. In *22nd International Conference On Computers in Education* (pp. 436–445).

Main, S., & O'Rourke, J. (2011). 'New Directions for Traditional Lessons': Can Handheld Game Consoles Enhance Mental Mathematics Skills? *Australian Journal of Teacher Education*, *36*(2). doi:10.14221/ajte.2011v36n2.4

Martin, L. (2018). *The Effect of Game-Based Learning on Title 1 Elementary Students' Math Achievement* (Ed.D. Dissertations, Concordia University, Portland).

Mavridis, A., Katmada, A., & Tsiatsos, T. (2017). Impact of online flexible games on students' attitude towards mathematics. *Educational Technology Research and Development*, *65*(6), 1451–1470. doi:10.1007/s11423-017-9522-5

McCue, C. M. (2011). *Learning middle school mathematics through student designed and constructed video games* (Theses, Dissertations, Professional Papers, and Capstones, University of Nevada, Las Vegas).

Miller, D. J., & Robertson, D. P. (2010). Using a games console in the primary classroom: Effects of 'Brain Training' programme on computation and self-esteem. *British Journal of Educational Technology*, *41*(2), 242–255. doi:10.1111/j.1467-8535.2008.00918.x

Miller, D. J., & Robertson, D. P. (2011). Educational benefits of using game consoles in a primary classroom: A randomised controlled trial: Game consoles in a primary classroom. *British Journal of Educational Technology*, *42*(5), 850–864. doi:10.1111/j.1467-8535.2010.01114.x

Pareto, L., Arvemo, T., Dahl, Y., Haake, M., & Gulz, A. (2011). A Teachable-Agent Arithmetic Game's Effects on Mathematics Understanding, Attitude and Self-efficacy. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Artificial Intelligence in Education* (Vol. 6738, pp. 247–255). doi:10.1007/978-3-642-21869-9_33

Riconscente, M. M. (2013). Results From a Controlled Study of the iPad Fractions Game Motion Math. *Games and Culture*, *8*(4), 186–214. doi:10.1177/1555412013496894

Rodríguez-Aflecht, G., Brezovszky, B., Pongsakdi, N., Jaakkola, T., Hannula-Sormunen, M. M., McMullen, J., & Lehtinen, E. (2015). Number navigation game (NNG): Experience and motivational effects. In J. Torbeyns, E. Lehtinen, & J. Elen (Eds.), *Describing and studying domain-specific serious games* (pp. 171–189). doi:10.1007/978-3-319-20276-1_11

Starkey, P. L. (2013). *The Effects of Digital Games on Middle School Students' Mathematical Achievement* (Theses and Dissertations, Lehigh University).

References

Begg, C. B., & Mazumdar, M. (1994). Operating Characteristics of a Rank Correlation Test for Publication Bias. *Biometrics*, *50*(4), 1088. doi:10.2307/2533446

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (Eds.). (2009). *Introduction to meta-analysis*. OCLC: ocn263294996. Chichester, U.K: John Wiley & Sons.

Cook, R. D. (1977). Detection of Influential Observation in Linear Regression. *Technometrics*, *19*(1), 15. doi:10.2307/1268249

Del Re, A. C., & Hoyt, W. T. [W. T.]. (2014). *MAd: Meta-analysis with mean differences*.

Duval, S., & Tweedie, R. (2000). Trim and Fill: A Simple Funnel-Plot-Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis. *Biometrics*, *56*(2), 455–463. doi:10.1111/j.0006-341X.2000.00455.x

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634. doi:10.1136/bmj.315.7109.629

Fisher, Z., & Tipton, E. (2015). Robumeta: An R-package for robust variance estimation in meta-analysis. *arXiv:1503.02220 [stat]*. arXiv: 1503.02220 [stat]

Fisher, Z., Tipton, E., & Zhipeng, H. (2017). *Robumeta: Robust variance meta-regression*.

Fox, J. (1991). *Regression diagnostics vol. 79: An introduction*. SAGE Publications Incorporated.

Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd, pp. 357–376). New York: Russell Sage Foundation.

Hedges, L. V., & Olkin, I. (2014). *Statistical methods for meta-analysis*. Academic press.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. doi:10.1002/jrsm.5

Hoyt, W. T. [William T.], & Del Re, A. C. (2018). Effect size calculation in meta-analyses of psychotherapy outcome research. *Psychotherapy Research*, *28*(3), 379–388. doi:10.1080/10503307.2017.1405171

Hung, C.-M., Huang, I., & Hwang, G.-J. (2014). Effects of digital game-based learning on students' self-efficacy, motivation, anxiety, and achievements in learning mathematics. *Journal of Computers in Education*, *1*(2-3), 151–166. doi:10.1007/s40692-014-0008-8

Ke, F. (2006). *Computer-based game playing within alternative classroom goal structures on fifth-graders' math learning outcomes: Cognitive, metacognitive, and affective evaluation and interpretation* (Doctoral dissertation, The Pennsylvania State University).

Ke, F. (2008). Alternative goal structures for computer game-based learning. *International Journal of Computer-Supported Collaborative Learning*, *3*(4), 429–445. doi:10.1007/s11412-008-9048-2

Landau, W. M. (2018). The drake R package: A pipeline toolkit for reproducibility and high-performance computing. *Journal of Open Source Software*, *3*(21).

Lin, L., & Chu, H. (2018). Quantifying publication bias in meta-analysis: Quantifying Publication Bias. *Biometrics*, *74*(3), 785–794. doi:10.1111/biom.12817

Moeyaert, M., Ugille, M., Natasha Beretvas, S., Ferron, J., Bunuan, R., & Van den Noortgate, W. (2017). Methods for dealing with multiple outcomes in meta-analysis : a comparison between averaging effect sizes, robust variance estimation and multilevel meta-analysis. *International Journal of Social Research Methodology*, *20*(6), 559–572. doi:10.1080/13645579.2016.1252189

Morris, S. B. (2008). Estimating Effect Sizes From Pretest-Posttest-Control Group Designs. *Organizational Research Methods*, *11*(2), 364–386. doi:10.1177/1094428106291059

Pareto, L., Arvemo, T., Dahl, Y., Haake, M., & Gulz, A. (2011). A Teachable-Agent Arithmetic Game's Effects on Mathematics Understanding, Attitude and Self-efficacy. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Artificial Intelligence in Education* (Vol. 6738, pp. 247–255). doi:10.1007/978-3-642-21869-9_33

Pigott, T. D., & Polanin, J. R. (2019). Methodological Guidance Papers: High-Quality Meta-Analysis in a Systematic Review. *Review of Educational Research*, 23. doi:10.3102/0034654319877153

Pustejovsky, J. (2019). *clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections*.

R Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/. Vienna, Austria.

Rodríguez-Aflecht, G., Brezovszky, B., Pongsakdi, N., Jaakkola, T., Hannula-Sormunen, M. M., McMullen, J., & Lehtinen, E. (2015). Number navigation game (NNG): Experience and motivational effects. In J. Torbeyns, E. Lehtinen, & J. Elen (Eds.), *Describing and studying domain-specific serious games* (pp. 171–189). doi:10.1007/978-3-319-20276-1_11

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Wiley.

Tipton, E. (2015). Small sample adjustments for robust variance estimation with Meta-Regression. *Psychological Methods*, *20*(3), 375–393. doi:10.1037/met0000011

Ushey, K. (2019). *Renv: Project environments*.

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2015). Meta-analysis of multiple outcomes: A multilevel approach. *Behavior Research Methods*, *47*(4), 1274–1294. doi:10.3758/s13428-014-0527-2

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

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, *1*(2), 112–125. doi:10.1002/jrsm.11