Multilevel Modeling
Multilevel Approach to Meta-Analysis

Youhee Kil (r0768512)

Spring 2020

Contents

1 Introduction **3**

2 Data and Research Questions **4**

2.1 Data 4

2.2 Research Questions 4

3 Methods and Analyses **5**

3.1 Multilevel Approach to Meta-Analysis 5

3.2 (a) Two-Level Model (Fixed Effect Model & Random Effect Model) 7

3.2.1 Result & Interpretation 8

3.3 (b) Three-Level 10

3.3.1 Result & Interpretation 10

4 Bayesian Hierarchical Model applying meta-analysis **11**

4.1 Background of Bayesian Hierarchical Modeling 11

4.2 Bayesian Hierarchical Model in Meta-Analysis 12

5 Conclusions **15**

6 Appendix **16**

References **18**

1 Introduction

It is relatively unknown to apply the multilevel method to meta-analysis, but it works well when we need to deal with the dependency of effect sizes (Assink et al. (2016)). Hox et al. (2010) stated that, in the past, only specialists used to perform multilevel modeling, but in these days, multilevel modeling approach can be easily seen in different fields where consisting of hierarchical structures, such as longitudinal data, sociometric modeling, and meta analysis. However, the expression of *multilevel model* is perplexing, because there are lots of different model names for hierarchical structured data, for example, multilevel model, hierarchical model, random effects model, mixed effects model, empirical Bayes and so on. Gelman and Hill (2006) expressed that *multilevel model* is, "a generic term for all models for nested data". In addition, *multilevel analysis* is, "examining relations between variables measured at different levels of the multilevel data structure". Multilevel model is often called as a *hierarchical model*, according to Gelman and Hill (2006), there are two reasons why it called as hierarchical modeling, first, the structure of the data is hierarchical. Second, the model itself having hierarchy. The meta-analysis is applied to first case, the nature structure of the data for meta-analysis is two-level which is clearly hierarchical. As Pastor and Lazowski (2018) addressed, there are still remaining confusions of using multilevel in meta-analysis, such as two-level meta-analysis and three-level meta-analysis. Therefore, in this paper, we present two-level model and three-level model to emphasize the differences between the two different level of models with multilevel approaches in meta-analysis.

In three-level model, there are some benefits to use three-level model for meta-analysis which is described as the extensions of a mixed-effect meta-analytic model. It has not been commonly used but when we are using same sample, it is relatively easier to calculate the multiple effect sizes (Van den Noortgate et al. (2013).

Normally, Bayesian modeling offers more opportunities in hierarchical meta-analysis than a classical hierarchical approach Lesaffre and Lawson (2012). Therefore, Bayesian approaches to multilevel meta-analytic model is briefly presented in this paper as well.

In Bayesian hierarchical modeling applied to meta-analysis one does not assume that several parameters from all investigated studies are completely independent from each other. This idea had been implicitly introduced by the founding father of meta-analysis Gene Glass. As McElreath (2020) points out, Bayesian hierarchical modeling, regardless whether it is applied in meta-analysis or in any other research field, is not memoryless. In fact, one can learn about the parameter in one study, based on the parameter derived from another study. Furthermore, one can learn not only about the individual parameters of investigated studies, but also the parameters of the whole population of studies.

This paper contributed to the analyses of the data from Cooper et al. (2003) "The Effects of Modified School Calendars on Student Achievement and on School and Community Attitudes" by focusing specifically on Multilevel modeling approaches in meta-analysis. Furthermore, we attempted to analyze further by using Bayesian hierarchical meta-analysis.

The contents of the paper is organized as follows. Section 2 briefly describes about the data from Cooper et al. (2003) and discusses related research questions. Section 3 presents the methods and results, (a) two-level model (Fixed-effect model and Random-effect model) and (b) three-level model are fitted. Section 4 introduces Bayesian approaches to multilevel meta-analytic data in short as a way of overcoming the limitations of multilevel and meta-analysis. Lastly, short conclusion is followed.

2 Data and Research Questions

2.1 Data

The data from Cooper et al. (2003) "The Effects of Modified School Calendars on Student Achievement and on School and Community Attitudes" is used to conduct multilevel meta-analysis (Table 1). The aim of the study in this paper is to generate a new understanding by combining and comparing the results of multiple conducted studies on the effects of modified school calendars on student achievement (Cooper et al. (2003)). The definition of the terminology *modified school* is that "Children might go to school for 9 weeks and then have 3 weeks off, or go for 12 weeks and have 4 weeks off but did not increase the length of their school year" (Cooper et al. (2003)), hence school where having modified school calendar consisted of more frequent but momentary intermittent breaks. There are some arguments about benefits and drawbacks of Modified Calendars, therefore, we aim to investigate if there is an effect on Modified Calendars and if the effect exists, then how the effect differs by the district of the study.

Cooper et al. (2003) used the d as an effect size which is the standardized mean difference in achievement test scores. Hence, the estimation of the standardized mean difference is the effect of calendar variation on measures of student achievement (Cooper et al. (2003)). The data is consisted of 56 studies with 11 different schools and districts in 4 different years to perform multilevel meta-analysis and Bayesian hierarchical meta-analysis with software R *nlme, metafor, brms* packages, the more detailed explanation of collecting and selecting studies can be checked in the paper of Cooper et al. (2003).

	district	**school**	**study**	**year**	**yi**	**vi**
1	11	1	1	1976	-0.180	0.118
2	11	2	2	1976	-0.220	0.118
3	11	3	3	1976	0.230	0.144
4	11	4	4	1976	-0.300	0.144
5	12	1	5	1989	0.130	0.014
6	12	2	6	1989	-0.260	0.014
7	12	3	7	1989	0.190	0.015
8	12	4	8	1989	0.320	0.024
9	18	1	9	1994	0.450	0.023
10	18	2	10	1994	0.380	0.043

Table 1: Some of the dataset

2.2 Research Questions

A lot of things have changed but education system, especially the school calendars always remains the same. consistently there are some demands of the changes of the education system. As an example, there are movement of some schools trying to have different calendar system. Therefore, investigating the effects of modified school calendars on student achievement is getting more necessary. Especially this kinds of investigation will be helpful for policymakers who are in charged of the education sector of the country.

Based on the data, following research questions are stated: (1) Is there modified calendar effect on students' achievement? (2) Does the calendar effect on students' achievement differ by the district of the school?

In this case, the null hypothesis is that there is no calendar effect. Positive d effect sizes reflect the modified calendar student yielded better achievement result. On the other hand, negative d effect sizes indicates the traditional school calendar student got better achievement result.

3 Methods and Analyses

3.1 Multilevel Approach to Meta-Analysis

A meta-analysis is known for, "a systematic approach toward the synthesis of a large number of results from empirical studies." Hox (1995) described that the most important preliminary question in meta-analysis is, whether the results differ more than each other than corresponds to the random sampling variation that is expected given the studies' sample sizes. There are two important terminologies appeared, *homogeneous* and *heterogeneous*. Simply, *homogeneous* means that they come from a single population, and *heterogeneous* means that they come from a different population. For example, for *heterogeneous* case, the results are different more than expected, then in this case estimating the 'average' result is not the primary goal anymore, instead of analyzing the excess variation becomes a goal (Hox (1995)). Hence, Hox et al. (2010) described that the goal for meta-analysis is not only to summarize the findings of multiple independently conducted studies on a specific research problem but also to derive study characteristics. Analogous to multilevel modeling, in a meta-analysis, we have a transformed dataset with a number of statistics such as correlation, in this case, then we assess their average value and variability by using the data. Finally, draw conclusions based on this information.

Pastor and Lazowski (2018) argued that the term "multilevel meta-analysis" is redundant, because all meta-analysis can be viewed as a "special case of multilevel model". Admittedly, most of the dataset for meta-analysis is consisted of two-level hierarchical linear structures, for example, first-level subjects within the studies, and the second level is the studies. We can predict the outcome variable of the meta-analytic data by employing the available explanatory variables such as individual level (level 1) and study level (level 2) and so on. Flexibility is the significant benefit of using a multilevel approach to meta-analysis instead of the classical meta-analysis method (Hox et al. (2010)). For example, adding more levels to explain the data further is not a big problem in multilevel meta-analysis. In addition, multilevel models offer an explicit framework in which to express similarity (exchangeability) judgments in order to combine information across units to produce accurate and well-calibrated predictions of observable outcomes (Draper (1995)).

Then, why meta-analysis is assumed nature hierarchical structures.

Firstly, the **random effects model** we have known is

$$\hat{\theta}_i = \mu + \epsilon_i + \zeta_i \tag{1}$$

assuming ϵ_i and ζ_i are two sources of variability, ϵ_i is a sampling error of individual studies and ζ_i is the between - study heterogeneity. Therefore, the target is to estimate the mean of the distribution of true effect sizes, μ, in the random-effects model. In meta-analysis data, there are also two levels corresponding those two sources of variability. The 'participant' level as level 1 and 'study' level as level 2. Therefore, it is possible to call these nested data, likewise the participants are nested within studies.

	Fixed Effects model	Random Effects model
Level 1	$\hat{\theta}_i = \theta_i + \epsilon_i$	$\hat{\theta}_i = \theta_i + \epsilon_i$
Level 2	$\theta_i = \mu$	$\theta_i = \mu + \zeta_i$
Combined	$\hat{\theta}_i = \mu + \epsilon_i$	$\hat{\theta}_i = \mu + \zeta_i + \epsilon_i$

Table 2: FE and RE models

Choice	Traditional	Multilevel
Estimation of τ^2	DerSimonian and Laird (DL)	Maximum Likelihood (REML, ML)
Hypothesis Significance Test	Q-statistics	Likelihood ratio tests
Analytical Path	Fixed-Effects (FE) Model	Random effects (RE) Model

Table 3: Differences between traditional and multilevel meta-analysis (Pastor and Lazowski (2018))

Level 1 (participants) model:

$$\hat{\theta}_i = \theta_i + \epsilon_i \tag{2}$$

Level 2 (study) model:

$$\theta_i = \mu + \zeta_i \tag{3}$$

Even a fixed effects model can be formed as multilevel structures with $\zeta_i = 0$. Therefore, it is clear to argue that meta-analytic model already has hierarchical structures given that participants are nested within studies in meta-analysis data. The summary of the Fixed-Effect model and Random-Effect model is presented in Table 2.

According to Hox et al. (2010), classical meta-analysis has devoted to develop system to integrate the variety of methods of statistics into one overall outcome and test these outcomes if it is regarded as homogeneous or heterogeneous. Then what is the difference between classical meta-analysis and multilevel meta-analysis? We ignored the moderator in this case since our model does not include moderator, hence only the difference is whether the model is a fixed-effects (FE) or a random-effects (RE). The RE model in Table 1 is clearly multilevel model since multilevel model have more than one random effect (ζ_i, ϵ_i), at the same time, actually the FE can be considered multilevel. In fact, the meta-analytic choices being made is the noteworthy difference between classical meta-analysis and multilevel meta-analysis rather than the models being used. In multilevel meta-analysis, maximum likelihood such as REML, ML is used to estimate τ and likelihood ratio test are used for a hypothesis significance test.

A commonly used estimator in Multilevel meta-analysis to estimate the between-study variances (τ^2) is maximum likelihood, restricted maximum likelihood (REML) and full maximum likelihood (ML) (Table 3). Pastor and Lazowski (2018) addressed that REML is preferred to use compared to ML in multilevel modeling and meta-analysis. But, it is not always true, therefore, different estimator has to be used to perform meta-analysis beforehand to consider which estimator will give a best result (Pastor and Lazowski (2018)).

When maximum likelihood estimation is used to estimate, Likelihood ratio test is used as a null hypothesis significance test in multilevel approach to meta-analysis. The approaches to fit the data is different between traditional and multilevel approaches, that's why Pastor and Lazowski (2018) came up with the terminology of "analytic path" to distinguish the path depends on the approaches. Fixed-Effect model is mostly placed in traditional approach, because fixed-model was used more frequently throughout the history, on the other hand, Random-Effect model is always begun in multilevel modeling.

Then discussion of the difference between the ordinary multilevel analysis and a typical meta-analysis is needed at this point. According to Van den Noortgate et al. (2013), there are three major differences,

(1) meta-analysis often used the data which combined results of the studies often with different scale. (2) meta-analysis often depends on the summary statistics, such as test statistics, effect size values, on the other hand, ordinary multilevel analyses often use with the raw data. (3) the most commonly used effect size metrics is different between the raw data multilevel analyses and meta-analyses. The remark of these differences is that multilevel model is relevant to perform in meta-analysis as long as the sampling distributions are approximately normal with an estimable variance, and all studies have used the same effect size metric (Van den Noortgate et al. (2013)).

3.2 (a) Two-Level Model (Fixed Effect Model & Random Effect Model)

In this case, standardized mean differences is used as an effects size, we divided the differences between the traditional school calendar mean from the modified calendar school on student achievement mean by the average standard deviation.

Estimating the standardized mean difference from the two independent groups, the traditional calendar and the modified calendar as

$$d = (\frac{\bar{X}_1 - \bar{X}_2}{S_{within}}) \tag{4}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the sample means in the two groups, S_{within} is the standard deviation of the within-group. By pooling the two estimates of the standard deviation,

$$S_{within} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_1^2}{n_1 + n_2 - 2}} \tag{5}$$

where n_1, n_2 are the sample sizes and S_1, S_2 are the standard deviation in the two groups, it yields a more accurate estimate of their common value (Borenstein et al. (2011)).

The variance was computed as

$$V_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)} \tag{6}$$

The column yi contains the effect size (d - 'standardized mean difference' in this case), the vi column contains the sampling variance.

As we indicated from the theory section, here is random-effect model:

$$y_i \sim N(\theta_i, \sigma^2)$$
$$\theta_i \sim N(\mu, \tau^2)$$

We usually assume that true effects are normally distributed in a random-effects meta analysis. Therefore, in this case, y_i is a draw from normal distribution which is centered on that study's effect size θ_i and has standard deviation equal to the study's observed standard error σ_i (Vuorre (2016)). In addition, the parameters θ_i are drawn from a normal distribution with hyperparameters (μ, τ) in this case. Gelman stated that θ_i's are conditionally independent given (μ, τ) and hierarchical model also permits the interpretation of the θ_i's as a random sample from a shared population distribution (Gelman et al. (2014))

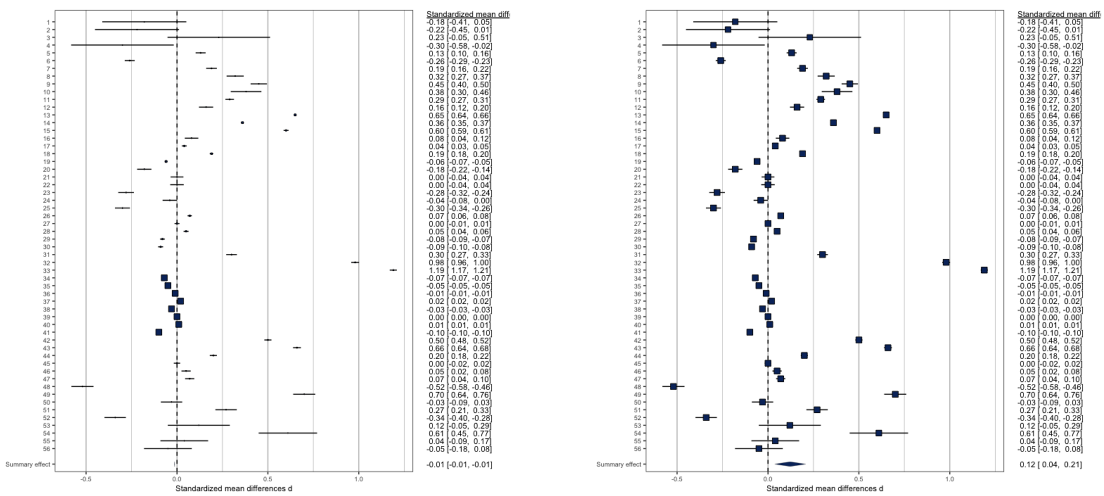
There are several packages in R to conduct random effect meta-analysis, such as *metafor* and *nlme*. In this case we mainly used *metafor* packages. Following mathematical formula for the model fitted by *rma()* function is referred from Viechtbauer (2018) and the package *metafor* (Viechtbauer (2010)).

$$
\begin{aligned}
y_i &= \mu + u_i + e_i \\
u_i &\sim N(0, \tau^2) \\
e_i &\sim N(0, v_i)
\end{aligned}
$$

where the v_i values are the known sampling variances. The model from *rma()* works better for this case.

3.2.1 Result & Interpretation

Figure 1 is forest plots of two models, fixed effects (FE) model and random effects (RE) model, presenting the standardized mean differences between the traditional school calendar and the modified calendar school on student achievement mean. The average standardized mean differences across the 56 studies in the fixed effect model was d = -0.01 (p < 0.001) with [-0.01, -0.01] (95% Confidence Interval) and in the random effect model with Maximum Likelihood was d = 0.12 (p = 0.003) with [0.04, 0.21] (95% Confidence Interval). It is clear to be seen that positive d in Random-Effect Model and slightly negative d in Fixed-Effect model. But confidence interval of these models does not include $d=0$, hence, In terms of hypothesis testing, the null hypothesis is rejected meaning that there is an calendar effect in this case. Each model's results are presented in Table 4, for example, z-value and p-value in RE are 2.945 and 0.003 respectively, hence, the null hypothesis can be rejected. However, a z-value and p-value is not an effect size, the aim of approach of p-value and effect size is different. It is a matter of choice which one to report, because two approaches are altered by the needs of the researcher such as p-value is for the viability of the null hypothesis and effect size is for the magnitude of the effect (Borenstein et al. (2011)). In Figure 1, one of the remarkable differences is the size of the bullet representing a study's weight. Therefore, there is relatively heavier weights on Random-Effects model.



(a) Fixed Effects Model (b) Random Effects Model

Figure 1: Forest plots

Model	estimate	se	zval	pval	ci.lb	ci.ub	
FE	0.046	0.009	5.050	<.001	0.028	0.064	***
RE (ML)	0.128	0.043	2.945	0.003	0.043	0.213	**
RE (ML- rma.mv)	0.128	0.044	2.916	0.004	0.042	0.214	**
Three-level	0.184	0.080	2.292	0.022	0.027	0.342	*

Table 4: Camparison of Model Results

Heterogeneity Prediction Interval

In addition, heterogeneity can be assessed easily by checking forest plot. In the plots, few of individual studies are not passing through overall effect indicating the existence of heterogeneity in the data. Cochran's Q test and index of heterogeneity I2 (I-squared) were tested to check heterogeneity. According to the result, p-value of both models (FE, RE) is p <0.05 suggesting heterogeneity. In addition, index of heterogeneity (I2) of FE is 90.50% and RE is 94.59%. Both I2 lies between 75% and 100% suggesting considerable heterogeneity. It validated us to use 'Random Effect' Model rather than 'Fixed effect' Model. The assumption of RE is that there is heterogeneity. As we stated when there is heterogeneous in data, then our goal becomes to analyze the dispersion of effect sizes. Interpreting the combined effects as meaningful and conducting significance test are not really acceptable when I2 is higher than 25% (Hak et al. (2016)). The dispersion and distribution of true effect sizes is assessed by the prediction interval (Borenstein et al. (2011)). In forest plot of RE, prediction interval is presented at the bottom of the plot. The interpretation of the prediction interval in meta-analysis is different as other primary studies. The prediction interval is suggested to interpret as a description of the range of observed effect size (Hak et al. (2016)).

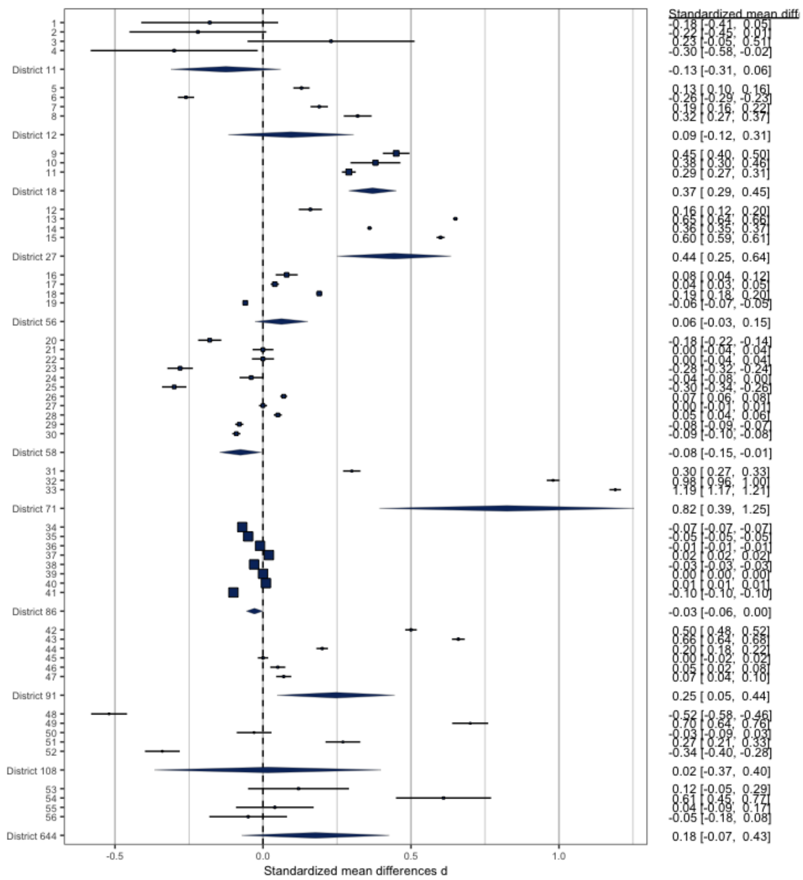


Figure 2: Forest Plot on the subgroup (District) Analysis

3.3 (b) Three-Level

As we have seen earlier, meta-analysis has naturally hierarchical structure, therefore, describing two-level model as multilevel meta-analysis is redundant. But still there are some researchers described the standard two-level model as a multilevel such as they listed two-level model as a three-level meta-analysis which is incorrect (Pastor and Lazowski (2018)). Hence, in this section, we are going to present three-level model to show the difference compared to the two-level model. Because, three-level meta-analysis model allowed us to analyze the data better, as we mentioned in the introduction section, Assink et al. (2016) stated that multilevel model is one of powerful ways to consider the dependency of effect sizes in meta-analysis.

Multilevel model helps to induce dependence in the data such as study from the same district is more alike compared to study from different school (Van den Noortgate et al. (2013)).

The best advantage of performing multilevel approach to meta-analysis is *Flexibility.* In multilevel meta-analysis, adding one more level for a better reflection of the data can be done simply. Simply, the three-level Model allows effect sizes to vary between observations (level 1), study (level 2), and district (level 3) in this case. The formulation is referred it from Harrer et al. (2019):

Level 1

$$\hat{\theta}_{ij} = \theta_{ij} = \epsilon_{ij} \tag{7}$$

Level 2

$$\theta_{ij} = k_j + \zeta_{(2)ij} \tag{8}$$

Level 3

$$k_j = \beta_0 + \zeta_{(3)j} \tag{9}$$

Altogether with Level 1, Level 2, Level 3 can be formulated as:

$$\hat{\theta}_{ij} = \beta_0 + \zeta_{(2)ij} + \zeta_{(3)j} + \epsilon_{ij} \tag{10}$$

where $\hat{\theta}_{ij}$ is the estimator of true effect size in ith effect size in cluster j, k_j the average effect size in j and B_0 the average population effect (Harrer et al. (2019)).

3.3.1 Result & Interpretation

A three-level random effect models applying meta-analysis is performed by using *rma.mv* function in R. The function *rma.mv* is a flexible way of fitting a multivariate/multilevel fixed- and random.mixed-effects models with or without moderator to meta analysis (Viechtbauer (2016)). There is a clear evidence for heterogeneity in this case by looking at forest plot, hence moderator analyses can be conducted. The purpose of moderator analyses is to determine whether there is a relationship between study and individual variables depends on the value of a district variable. In addition, Assink et al. (2016) stated moderator analyses is used to test to account for heterogeneity for within-study or between-study of variables. The forest plot of subgroup analysis is presented and district is treated as a moderator/subgroup in Figure 2. For example, the effect size of the group 'district 11' is -0.13 with the confidence interval [-0.31, 0.06] at 95%, on the other hand, the effect size of the group 'district 71' is 0.82 with the confidence interval [0.39, 1.25] at 95%. However, moderator analyses is not conducted instead three-level meta analysis is performed with another level from two-level meta-analysis in this case. The details of the outcomes of three-level multilevel model in meta analysis is discussed in following:

- **K=56; method = "ML"** : k is indicating the number of effect sizes, hence, 56 effect sizes are used with Maximum Likelihood estimation method (ML). There are several more popular estimation methods in multilevel meta-analysis. The most common one is the Restricted Maximum likelihood estimation method (REML), some researchers said the REML method is better than other estimation methods, but it has some restrictions as well, and sometime it yields almost similar results as Maximum Likelihood.
- **Variance Component**: The variance of the second level of the model is 0.058 which is for the variance between effect sizes within studies (σ_1^2) and the variance of the third level of the model is 0.033 which is for the variance between studies (σ_2^2).
- **Test for Heterogeneity** : The p-value for heterogeneity is $p < .001$, but the result of heterogeneity is not a big concern, investigating within-study variance (level 2) and between-study (level 3) is more interesting.
- **Model Results**: The model result of three-level is presented in Table 4. the overall mean standardized difference is 0.184 with a standard error of 0.080, and the confidence interval is [0.027, 0.342]. According to the Cohen (1988), the criterion of effect sizes indicates that *d=0.2* is small effects. Based on the result of Table 4, it indicates that the effect size is generally small effects. Therefore, even though we can detect some positiveness of the effect of modified calendar, it is very mild. However, based on the Figure 2, the different average effect sizes are detected by district, investigating each district will be required for the further investigation.

Intraclass Corrletaion Coefficient of True effect

The indication of intra-class correlation (ICC) is the degree of between-study variation on the subject of the total variance (Cheung (2014)). If the effect sizes are correlated which suggests that there is a dependency between effect sizes, then this situation will decrease the heterogeneity leading to false-positive results (Harrer et al. (2019)). But, referred to the Cheung (2014), the three-level model in the meta-analysis allows the effect sizes to be correlated within *districts*. If there is no dependence, the result of ICC equals to 0. Therefore, if the ICC is 0 at level 3, it indicates the three-level model is same as two-level model, because it means there is no variation in between-study (level3) since ICCs in level 2 and level 3 suggests the variation at each level (Cheung (2014)). The formulation of ICC is referred from:

$$p = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \tag{11}$$

where variance component σ_1^2 is corresponding to the *study* variables and σ_2^2 is corresponding to *district.*

In this case, the result of ICC is 0.665 indicating there is somewhat strong dependency within *district.*

4 Bayesian Hierarchical Model applying meta-analysis

4.1 Background of Bayesian Hierarchical Modeling

In this section, Bayesian Multilevel model applying meta-analysis is briefly introduced with *brms* packages. There are diverse sampling procedures for Bayesian, for example, Gibbs Sampling, Metropolis-Hastings Sampling etc. In this case, NO-U-Turn Sampling is used, that is the main sampling method in *brms* package. Bayesian approach will provide more reasonable inferences when the case doesn't allow to get a enough information from the model and data to estimate variance parameters such as the number of groups is small or the multilevel model is complicate (Gelman et al. (2014)).

The advantage of performing Bayesian Hierarchical Meta-Analysis is discussed in Harrer et al. (2019). First, when the number of studies is small, the uncertainty regarding τ^2 can be directly modeled with Bayesian methods and its better in estimating variance τ^2 (the between-study heterogeneity) and pooled effect. Second, credible interval which is the most practical way to interpret the interval. Third, the posterior distributions derived by Bayesian hierarchical Meta-Analysis is the integration of prior knowledge and likelihood (assumptions).

In Bayesian approach, it is essential to assume that each θ_j is dependent on another θ_j. Therefore, from Bayesian perspective, Bayesian hierarchical model is that the estimates of individual parameters are communicated by the data from all other items (Gelman et al. (2014)). This leads the parameters to depend on each other.

In addition, the concept of exchangeability is the most crucial concept for the Bayesian hierarchical model applying to the Meta-Analysis. Generally the concept of the exchangeability is "an infinite sequence of random quantities is exchanageable if any finite sub-sequence is exchangeable" (Migon et al. (2014)). By exchageable distribution we consider the distribution that assumes that each θ_j is independent identically distributed sample (Gelman et al. (2014)), to briefly summarize the concept of exchageability is followed :

$$p(\theta|\phi) = \prod_{j=1}^{J} p(\theta_j|\phi) \tag{12}$$

As we consider the following formulation of Bayesian hierarchical model (Gelman and Hill (2006)):

- a set of $j = 1, ..., J$;
- each j corresponds to observed data y_j;
- each j probability is described by θ_j;
- likelihood is given by $p(y_i|\theta_j)$

Now we are able to adapt these formulation of Bayesian hierarchical model to the context of meta-analysis:

- a set of studies $j = 1, ..., J$;
- each study, j has a study-specific number of observed y_j;
- each j probability is described by θ_j;
- likelihood function is given by $p(y_i|\theta_j)$

4.2 Bayesian Hierarchical Model in Meta-Analysis

The idea behind of Bayesian Meta-Analysis is referred from Harrer et al. (2019). In level 1, we have individual participants assuming that the observed effect size $\hat{\theta}_i$ of each study i. In level 2, we have study which can be assumed as participants from level 1 are nested within studies, and that θ_i follow

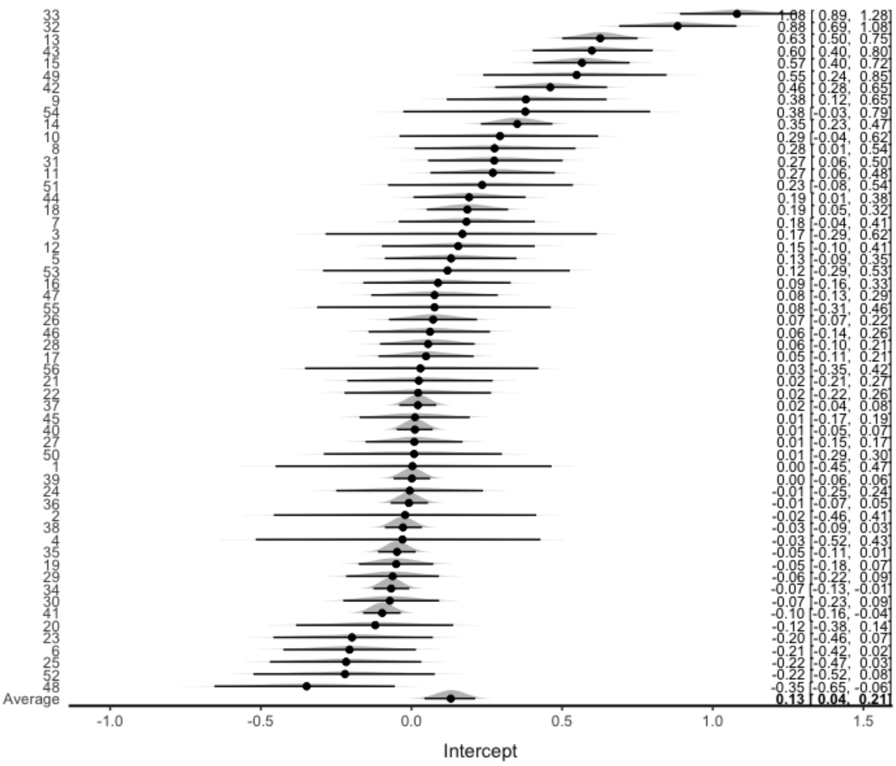


Figure 3: Forest Plot of Bayesian hierarchical two-level model

the normal distribution with a mean μ and variance τ^2 (the between-study heterogeneity). Therefore, the formula of Bayesian multilevel modeling in meta-analytic notation is :

$$
\begin{aligned}
y_i &\sim N(\theta_i, \sigma_i^2) \\
\theta_i &\sim N(\mu, \tau^2) \\
(\mu, \tau^2) &\sim p(.), \tau^2 > 0
\end{aligned}
$$

which is similar to random-effect model but, there is a remarkable difference $(\mu, \tau^2) \sim p(.), \tau^2 > 0$. Assigning prior distribution (p(.)) for μ and τ^2 and restricting τ^2 to positive is the first step of implementing Bayesian.

For the prior distribution, we assigned a prior distribution to τ. We set μ and τ as

$$
\begin{aligned}
\mu &\sim N(0, 1) \\
\tau &\sim HalfCauchy(0, 0.2)
\end{aligned}
$$

Gelman et al. (2014) described that weakly informative prior which is a middle ground between a fully informative and uninformative (flat) prior distribution. A prior does not necessarily have to reflect historical or empirical data, but can be chosen based on desirable mathematical properties (Williams et al. (2018)). However, finding an optimal estimator for τ^2 is important for several reasons, further details of weakly information prior distribution can be found in paper titled 'Bayesian Meta-Analysis with Weakly Informative Prior Distributions' (Williams et al. (2018)). In this case, the reason why we assumed $\tau \sim$ Half Cauchy(0, 0.2) is that average effect size of all figures is approximately close to 0.2 and the hierarchical half-Cauchy model performs a great deal of shrinkage (Gelman et al. (2014)). The Bayesian approach provides a mechanism for incorporating prior knowledge into an analysis (Enders

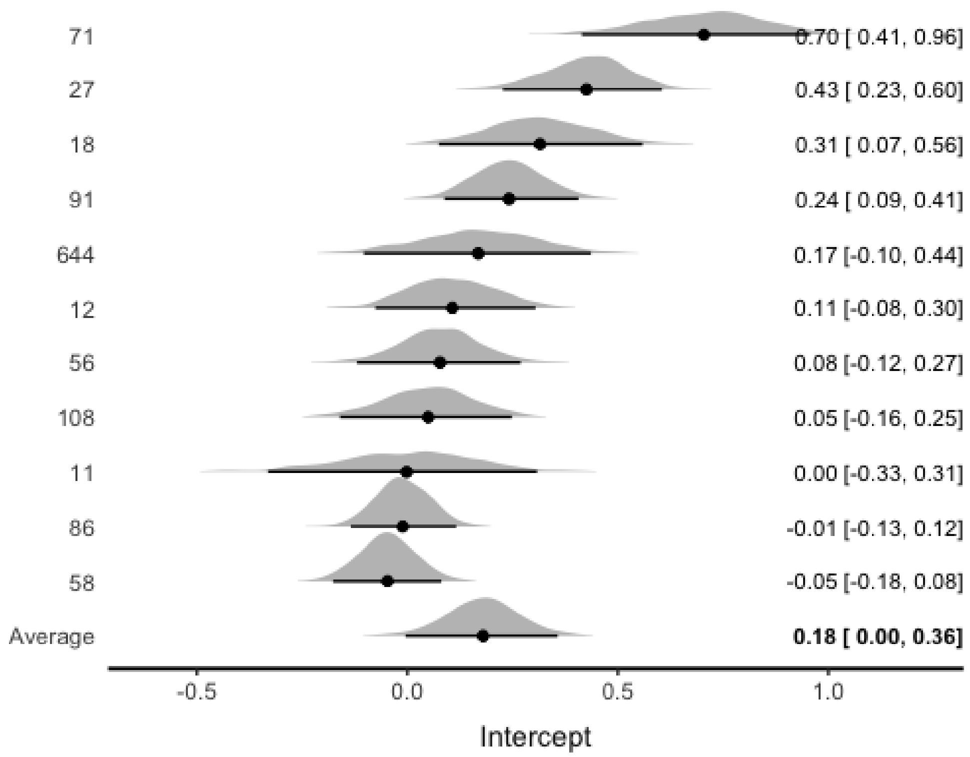


Figure 4: Forest Plot of Bayesian hierarchical three-level model

(2010)), therefore, in this case, we formulated a prior distribution by using information such as the average effect size from the meta-analysis.

Model	Estimate	Est.Error	l-95% CI	u-95% CI
Group-Level Effects:	0.30	0.03	0.24	0.37
Population-Level Effects:	0.13	0.04	0.21	1.01

Table 5: Results of Bayesian hierarchical models (two-level model)

Group-Level Effects:	Estimate	Est.Error	l-95% CI	u-95% CI
district (Number of levels: 11)	0.26	0.08	0.15	0.42
district:study (Number of levels: 56)	0.18	0.03	0.13	0.25
Population-Level Effects:	0.18	0.09	0.00	0.36

Table 6: Results of Bayesian hierarchical models (three-level model)

As Mila and Ngugi (2011) stated that one of benefit of performing Bayesian method on meta-analysis is to measure the uncertainty of the results easily by the 95% CrI. The Figure 3 displays each θ_i for posterior distribution. The mean and 95% CrI limits are displayed on the right side of the plot. In addition, the results of the Bayesian hierarchical meta-analysis are differed to Figure 1, when we compare the CrI and CI of each study and average effect size.

Table 5 represents Bayesian hierarchical model on two-level model with displayed group-level effects and population-level effects. By using the suggestion from Buerkner (2016), every parameter is summarized using the mean (‘Estimate’) and the standard deviation (‘Est.Error’) of the posterior distribution as well as two-sided 95% credible intervals (‘l-95% CI’ and ‘u-95% CI’). For hyper-parameters (population-level effects) the 95% credibility interval for the average size effect is between 0.21 and 1.01 To sum up, it means that we are 95% that the average size effect of calendar effect on students' achievement is between 0.21 and 1.01 for the whole population of studies. For individual parameters (group-level effects, θ_j for study j), we are 95% certain that the the average size effect of calendar effect on students' achievement is between 0.24 and 0.37.

Furthermore, Table 6 represents the Bayesian Hierarchical model on three-level model which displays the groups-level effects with two categories (district, district:study). The result of the Bayesian Hierarchical model on three-level is different compared to the Figure 2 (Forest Plot on the subgroup (district) Analysis, such as the average effect size remains the same but values of CrI and CI are different. Also, Figure 4 presents more readable to compare the effects of calendar on students' achievement by district. We are able to conclude clearly that District 71 has the biggest positive effect of a modified calendar on students' achievement. Generally it shows that there is a positive effect of the calendar, but when we check *District 58, 86, 11* are not having any effect of calendar on students' achievement.

Table 6 represents Bayesian hierarchical model on three-level model with displayed group-level effects (district, district:study) and population-level effects. For the hyper-parameters (population-level effects) the 95% credibility interval for the average size effect is between 0.00 and 0.36. To sum up, it means that we are 95% that the average size effect of calendar effect on students' achievement is between 0.00 and 0.36 for the whole population of studies. Understandably, the credibility Interval for district is wider than the studies by the district. For individual parameters (group-level effects, θ_j by district for study j), we are 95% certain that the average size effect of calendar effect on students' achievement is between 0.13 and 0.25. And, for the group-level effects for district, we are 95% certain that the average size effect of calendar effect on students' achievement is between 0.15 and 0.42. (All of Bayesian hierarchical model resulted the value of Rhat equals 1).

5 Conclusions

In this paper, multilevel modeling on Meta-Analysis is performed to answer the research questions: (1) Is there modified calendar effect on students' achievement? (2) Does the calendar effect on students' achievement differ by the district of the school? We also elucidated the theoretical and methodological aspects of multilevel meta-analysis and Bayesian approach to hierarchical modeling in meta-analysis.

In general, we attempted to analyze the data with diverse ways of multilevel models such as (1) Two-level model (Fixed-Effect model and Random-Effect model), (2) Three-level model, (3) Bayesian hierarchical model approach on well-established meta-analytic data via a package *metafor, brms* in R.

We detected a positive but little effect of modified calendar on students' achievement except fixed-effect model. But the calendar effect on students' achievement is differed by the district of the school. Therefore, further investigation is needed to seek if the economically low-income district has a greater effect of the modified calendar on students' achievement.

In conclusion, the Multilevel approach in meta-analysis offers more flexible computational and methodological opportunities as well as Bayesian hierarchical modeling in meta-analysis.

6 Appendix

```
library(Matrix)
library(metafor)
library(dplyr)
library(ggplot2)
library(brms)
library(metaviz)

df <- dat.konstantopoulos2011
data<- df %>%
  mutate(sei = sqrt(vi)) %>%
  select(district, school, study, year, yi, vi, sei)


res.FE <- rma(yi, vi, data=data, method ="FE")
res.RE <- rma(yi, vi, data=data, method ="ML")
res.EB <- rma(yi, vi, data=data, method ="EB")

print(res.FE, digits=3)
print(res.RE, digits=3)
print(res.EB, digits=3)

## Two level model (standard random effects model)

mafixed<- rma(yi, vi, data=data, method ="FE")
print(mafixed, digits=3)
ma<- rma(yi, vi, slab=study, data= data, method ="REML")
print(ma, digits=3)
forest(ma)
ml.ma <- rma.mv(yi, vi, random = ~1 | study, data = data)
print(ml.ma, digits =3)
three.ma <- rma.mv(yi, vi, random = ~1 | district/study, data = data)
print(three.ma, digits = 3)



par(mfrow=c(1,2))
viz_forest(x = data[1:56, c("yi", "vi")],
           study_labels = data[1:56, c("study")],
           summary_label = "Summary effect",
           xlab = "Standardized mean differences d", method="FE",
           annotate_CI = TRUE
)

viz_forest(x = data[1:56, c("yi", "vi")] ,
           study_labels = data[1:56, c("study")],
           summary_label = "Summary effect",
           xlab = "Standardized mean differences d", method="REML",
           annotate_CI = TRUE
)

viz_forest(x = data[1:56, c("yi", "vi")],
           group =data[1:56, "district"] ,
           study_labels = data[1:56, c("study")],
           xlab = "Standardized mean differences d",
           method="ML",
           summary_label = c("District 11", "District 12", "District 18", "District 27",
                             "District 56", "District 58", "District 71", "District 86",
                             "District 91", "District 108", "District 644"),
           annotate_CI = TRUE)


```

```r
#Three level model
threelevel <- rma.mv(yi, vi, random = list(~1 | district, ~1|study ), data = data,
    method ="ML")
summary(threelevel, digits =3)

round(threelevel$sigma2[1] / sum(threelevel$sigma2), 3)

library(dmetar)

#BHMA model

set.seed(1004)
get_prior( yi|se(sei) ~ 1 + (1|district/study), data = data)

prior_c <- c(set_prior("normal(0, 1)", class = "Intercept"),
             set_prior("cauchy(0, 0.2)", class = "sd"))

library(extraDistr)
phcauchy(0.3, sigma = 0.3)

get_prior(yi|se(sei) ~ 1 + (1|study), data = data)
brm1<- brm(
  yi|se(sei) ~ 1 + (1|study),
  prior = prior_c,
  data = data,
  cores = 2,
  file = NULL)

brm2<- brm(
  yi|se(sei) ~ 1 + (1|district/study),
  prior = prior_c,
  data = data,
  cores = 2,
  file = NULL)

summary(brm1)
library(brmstools)
#BHMA forest graph
forest(brm1)
forest(brm2)

#probability check
brm1 %>%
  plot(
    combo = c("hist", "trace"), widths = c(1, 1.5),
    theme = theme_classic(base_size = 10))


brm2 %>%
  plot(
    combo = c("hist", "trace"), widths = c(1, 1.5),
    theme = theme_classic(base_size = 10))
# investigate model fit
mcmc_plot(brm1)
pp_check(brm1)
mcmc_plot(brm2)
pp_check(brm2)
```

References

Assink, M., Wibbelink, C. J., et al. (2016). Fitting three-level meta-analytic models in r: A step-by-step tutorial. *The Quantitative Methods for Psychology*, 12(3):154–174.

Borenstein, M., Hedges, L. V., Higgins, J. P., and Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.

Buerkner, P. (2016). brms: bayesian regression models using stan. r package version 0.9. 0.

Cheung, M. W.-L. (2014). Modeling dependent effect sizes with three-level meta-analyses: a structural equation modeling approach. *Psychological Methods*, 19(2):211.

Cohen, J. (1988). Statistical power analysis for the social sciences.

Cooper, H., Valentine, J. C., Charlton, K., and Melson, A. (2003). The effects of modified school calendars on student achievement and on school and community attitudes. *Review of Educational Research*, 73(1):1–52.

Draper, D. (1995). Inference and hierarchical modeling in the social sciences. *Journal of Educational and Behavioral Statistics*, 20(2):115–147.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford press.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2014). *Bayesian data analysis (3d ed.)*. Boca Raton: Chapman and Hall/CRC Press.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Hak, T., van Rhee, H., and Suurmond, R. (2016). How to interpret results of meta-analysis. *Available at SSRN 3241367*.

Harrer, M., Cuijpers, P., Furukawa, T., and Ebert, D. (2019). Doing meta-analysis in r: A hands-on guide. *PROTECT Lab Erlangen*.

Hox, J. J. (1995). *Applied multilevel analysis*. TT-publikaties.

Hox, J. J., Moerbeek, M., and Van de Schoot, R. (2010). *Multilevel analysis: Techniques and applications*. Routledge.

Lesaffre, E. and Lawson, A. (2012). *Bayesian Biostatistics*.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.

Migon, H. S., Gamerman, D., and Louzada, F. (2014). *Statistical inference: an integrated approach*. CRC press.

Mila, A. and Ngugi, H. (2011). A bayesian approach to meta-analysis of plant pathology studies. *Phytopathology*, 101(1):42–51.

Pastor, D. A. and Lazowski, R. A. (2018). On the multilevel nature of meta-analysis: a tutorial, comparison of software programs, and discussion of analytic choices. *Multivariate Behavioral Research*, 53(1):74–89.

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., and Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior research methods*, 45(2):576–594.

Viechtbauer, W. (2010). Conducting meta-analyses in r with the metafor package. *Journal of statistical software*, 36(3):1–48.

Viechtbauer, W. (2016). Meta-analysis via multivariate/multilevel linear (mixed-effects) models.

Viechtbauer, W. (2018). A comparison of the rma.uni() and rma.mv() functions.

Vuorre, M. (2016). Meta-analysis is a special case of bayesian multilevel modeling.

Williams, D. R., Rast, P., and Bürkner, P.-C. (2018). Bayesian meta-analysis with weakly informative prior distributions.