How Prior Information from National Assessments can be used when Designing Experimental Studies without a Control Group

Don van den Bergh[1], Nina Vandermeulen[2], Marije Lesterhuis[2], Sven de Maeyer[2], Elke Van Steendam[3] & Gert Rijlaarsdam[1] & Huub Van den Bergh[4]

1 University of Amsterdam | The Netherlands
[2] University of Antwerp | Belgium
[3] KULeuven | Belgium | Belgium
[4] University of Utrecht | The Netherlands

Abstract: National assessments yield a description of the proficiency level in a domain while accounting for differences between tasks. For instance, in writing assessments the level of proficiency is typically evaluated with a variety of topics and multiple tasks. This enables generalizations from specific tasks to a domain. In (quasi-)experimental research, however, writing skills are often evaluated with a single task. Yet, conclusions about the effectiveness of the treatment are formulated on the level of the domain, which is, euphemistically put, quite a stretch. Although conclusions drawn about the effect of the treatment are specific to the task administered, they are often generalized to the domain without any form of reservation. This raises the question whether we can use the results of national assessments about differences between tasks in the analyses of experimental studies. In this paper, we demonstrate how the information of a baseline data set can be used as a kind of control condition in the analysis of an experimental study.

Keywords: Prior information, Baseline comparison, Bayesian inference

journal of WRITING RESEARCH

Van den Bergh, D., Vandermeulen, N., Lesterhuis, M., De Maeyer, S., Van Steendam, E., Rijlaarsdam, G., & Van den Bergh, H. (2023). How prior information from national assessments can be used when designing experimental studies without a control group. *Journal of Writing Research*, *14*(3), 447-469. DOI: 10.17239/jowr-2023.14.03.05

Contact: Don van den Bergh, University of Amsterdam, Department of Psychological Methods, Postbus 15906, 1001 NK Amsterdam | The Netherlands - d.vandenbergh@uva.nl.
Orcid: http://orcid.org/0000-0002-9838-7308

Copyright: This article is published under Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported license.

1. Introduction

In many countries, the achievements of students are monitored in so-called national assessments. For instance, NAEP in the US, PEIL in the Netherlands (or international assessment programs like IEA or PIRLS) measure students' achievements at regular intervals to gain information on changes in achievement over time (or changes in differences between countries). Although the results of these assessments often inform policymaking, the data are hardly ever used in educational research even though there are ample opportunities.

A common denominator in national assessments is that all subject domain measurements are based on a domain definition, based on an analysis of that domain. That is why, in the case of a writing assessment, students write multiple texts. This is a necessity if one wants to describe the level of achievement covering a whole domain while generalizing over specific assignments at the same time. For a writing assessment in the Netherlands, for instance, students took a sample of 3 out of 21 assignments (Zwarts et al., 1990) via an overlapping design to optimize the relation between testing time per individual while allowing for conclusions at the population level at the same time.

If we contrast experimental studies with national assessments it is apparent that in many experimental studies the measurements are not as varied as in national assessments. In the vast majority of experimental studies on writing, students write one text as a pretest and another text as a posttest (e.g., Graham & Harris, 2014). Based on these texts quite frequently conclusions are drawn about changes in students' writing skills. However, it is well documented that differences between different types of writing assignments can be large (e.g., Bouwer et al., 2015; Rijlaarsdam et al., 1992). Consequently, in an experimental study differences between pre and posttest can be interpreted as a task effect or as an intervention effect. Therefore, a control is required to show that a differential effect between measurements is not due to different tasks but rather due to the intervention. However, we can hardly make inferences based on only one writing assignment. Although many researchers are aware of the limited generalizability of single-task experiments, in most studies it is often infeasible that students write more tasks.

So, whereas we have got quite a lot of information on levels of achievement (at certain educational levels) on the basis of national assessments, for the majority of experimental studies, we must rely on relatively small samples of participants and relatively limited samples of domain specific tasks. Therefore, one could wonder why do we not use the information from large-scale assessments, in which students write multiple texts (or take many tests), in the analysis of experimental data? This might enhance the generalizability of the results of experimental studies.

In experimental studies, quite often the goal is to show that the increase in achievement due to an experimental manipulation exceeds 'natural' growth.

Usually, this is done by comparing an experimental condition to a control condition, where the control condition indicates the 'natural' growth of achievement. This is possibly inefficient, as students in the control group have to be measured repeatedly as well to obtain sufficient power. One alternative to increase the power of studies is by enriching the statistical analyses with prior results (Rijlaarsdam et al., 2012). In contrast to

control conditions, national assessments consist of more observations on a larger variety of tasks and therefore provide a much richer account of the general level of achievement. Thus, the idea is to use the data from a national assessment as prior knowledge on the general level of achievement instead of a control group.

Unfortunately, no straightforward method exists to incorporate prior information into analyses. Ideally, the raw data from prior studies is included in the analyses as a benchmark comparison, but this is often impossible for practical (and privacy) reasons. Alternatively, prior knowledge can be represented by treating the prior results as population values and experimental results can be tested against these values. However, this approach seems far from ideal, as measurement error and uncertainty in the prior results are ignored. Consequently, standard errors are underestimated which increases type-I errors when comparing prior results and experimental data.

In this paper we discuss a method to compare baseline data and experimental data while taking the uncertainty in both data sets into account. We adopt a Bayesian approach to quantify the uncertainty in both the analysis of the baseline data and experimental data. Consequently, we can express the effectiveness of the intervention in the experimental data relative to the baseline data and make inferences about the significance of the intervention.

The outline of this paper is as follows. First, we introduce a large data set on writing instruction in high school that serves as a baseline data set in which students wrote multiple texts. We analyze the baseline data set by means of a multilevel model. Next, we introduce a follow-up data set from an experiment in which students wrote a single text on three different occasions. The follow-up data set is analyzed with a multilevel model as well and we relate the parameter estimates to those of the baseline analysis. We conclude by discussing the applicability, benefits, and limitations of our approach.

2. Baseline Data Set

The baseline data set was collected to investigate the writing quality of students in the tenth, eleventh, and twelfth grade of high school. Here we provide some information about the data collection and some descriptives of the data.

Schools were selected at random by creating three lists of schools. First, a school in the first batch was approached for participating in the study. If this school did not reply or refused to participate, a school in the second batch was selected at random.

If the second school did not participate, a school from the third batch was approached.

In total, the writing quality was measured for 625 students, nested in 43 schools. To assess between-task variance, 32 different tasks were administered of which each students wrote four. Not all students made all tasks, 497 students made four tasks and 128 students made three or fewer tasks. The minimum number of students per task was 62 whereas the maximum was 84. A benchmark rating procedure was used to assess the students' texts. This procedure entails that texts are rated holistically by comparing them to five benchmark texts at intervals of 1 SD. Benchmark rating proved to be a reliable rating method in several previous studies on writing (Blok, 1986; Bouwer et al., 2017; De Smedt et al., 2016; Rietdijk et al., 2017).

An overlapping rater team design with a total of 48 raters was applied (Van den Bergh & Eiting, 1989). Every text was rated by a jury of three raters; average jury reliability was 0.65. The text's final score consisted of the average of the three separate scores.

The data have a clear nested structure: observations are nested within students and tasks, and students are nested within schools. Students took a sample of four tasks out of 32 tasks developed for this writing assessment. Figure 1 shows the observed differences between grades, tasks, and students. Scores of students in the same school might be more alike than scores of students from different schools. Likewise, scores on the same task might be more alike than scores on different writing tasks. Therefore, a cross-classified multilevel model is in operation.

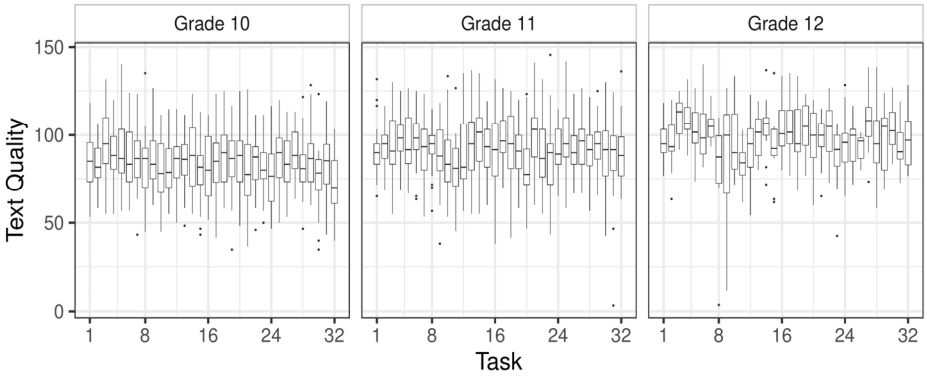


Figure 1. Box and whiskers plot of student performance on each task for the three grades measured. The grade is indicated above each panel and the task code is shown on the x-axis. There is substantial variance in student performance between tasks and within tasks. Student performance appears to increase in successive grades.

If $y_{(ij)k}$ is the score of student i ($i = 1, 2, 3, ..., I_k$) on task j ($j = 1, 2, ..., J_i$) in school k ($k = 1, 2, , K$), we can write the model that we will analyze as a function of student, task, school, and grade:

$$y_{(ij)k} = \beta_0 + \beta_1 \times [\text{Grade}_{(ij)k} = 11] + \beta_2 \times [\text{Grade}_{(ij)k} = 12] + w00k + u0k + v0j0 + \epsilon(ij)k.$$

The model consists of a fixed part, the first line, and a random part, the second line. In the fixed part, Grade(ij)k is an indicator matrix for students' grade (e.g., for students in grade 11 [Grade(ij)k = 11] = 1, otherwise 0). The intercept β0 represents the mean writing score in Grade 10. Consequently, the regression weights β1 and β_2 represent the difference in mean writing score between grade 11 and 10, and grade 12 and 10 respectively. In the random part four residual scores are distinguished, all of which are assumed to be normally distributed with an expected value of 0. The first residual (w_{00k}) captures the difference between the average of a school and the intercept. The second residual (u_{0k}) captures that the average score of student i in school k can deviate from the schools' average. The third residual (v_{0j0}) captures differences in difficulty between tasks. The fourth residual ($\epsilon_{(ij)k}$) indicates the deviation of the score of task j of the average of student i in school k. Usually, the variance of this term is interpreted as random noise.

To analyze the model we used a Bayesian approach. This decision is mostly a pragmatic one, as Bayesian inference is naturally accompanied by uncertainty estimates. For an introduction to Bayesian inference, see the special issue in *Psychonomic Bulletin & Review* which provides tutorials and guidance for aspiring Bayesians (Vandekerckhove et al., 2018).

3. Statistical Software

All analyses were conducted in R (R Core Team, 2022). The R package brms was used for Bayesian multilevel analyses (Bürkner, 2017). The R package brms is a convenient front-end for the probabilistic programming language Stan, which is software for general-purpose Bayesian inference (Carpenter et al., 2017). R code for all analyses is available at https://github.com/vandenman/Priors-Education.[1] For all analyses, we used six MCMC chains to assess convergence. Convergence was assessed using the $\hat{R}$ statistic (Vehtari et al., 2021). In line with the recommendations by Vehtari et al. (2021), we tweaked the parameters of the Stan algorithm such that the $\hat{R}$ is less than

[1] Although the original data cannot be shared for privacy considerations, the code in the GitHub repository contains a simulated dataset that can be used to replicate all the steps performed here. Furthermore, the MCMC samples are stored and can be used to replicate all follow-up analyses (e.g., recreate figures and tables).

1.01 and the rank-normalized effective sample size is larger than 400. Per chain, we simulated 60,000 samples and discarded the first 10,000 as warmup samples. In total, results in Tables and Figures are based on 300,000 samples of the posterior distribution.

We used the default prior distributions of the R package brms for all parameters. That is, the standard deviations of the random effects and the residual were assigned a half t-distribution with a mean of 0, scale of 18, and 3 degrees of freedom. For the fixed effects we used a Cauchy distribution with location 0 and scale 1. These priors are reasonably uninformative (Bürkner, 2017; Gelman, 2006) and given the sample size at hand, the influence of the prior is negligible.

4. Baseline Analysis

We summarized the posterior distribution in Table 1. This shows that the average text quality of students in grade 10 is estimated at 84.40. The 95% highest posterior density (HPD) credible interval ranges from 95% HPD [82.13, 86.67]. Students in grade 11 performed on average about 7.67 points better (95% HPD [5.50, 9.82]) than students in grade 10. Likewise, students in grade 12 performed on average about 13.46 points better (95% HPD [10.15, 16.79]) than students in grade 10. The estimated variance between schools (13.81), students within school (97.14), and tasks (10.89) clearly deviates from 0. Since the data set contained such a large variety of schools and tasks, these findings likely generalize over tasks.

Figure 2 visualizes the improvement in text quality across grades. To obtain the posterior distributions for grades 11 and 12, we add the posterior distribution of the intercept to that of the improvement in Grade 11 and Grade 12.

The descriptive plot in Figure 1 suggested that the scores of students depend on the task they made. This is further confirmed by the posterior mean of between task variance 10.89 (95% HPD [4.37, 18.84])) which is far away from 0. This shows that differences between tasks are substantial and that some tasks are systematically more difficult than others. Consequentially, a students' domain score depends on the task they made.

5. Application to an Experimental Analysis

In this section, we demonstrate how the findings based in the baseline analyses can be used as prior information to assess the impact of an intervention in an experimental analysis.

Table 1. Summary of the posterior distribution for the baseline data set. The first column shows the parameter, the second the posterior mean for that parameter, the third the posterior standard deviation and the last two columns show the 95% higher posterior density interval.

Grade 11 and 12 represent the improvement relative to grade 10 (the intercept). The posterior standard deviation may be interpreted as a standard error.

			95% HPD	
Parameter	*Mean*	*SD*	*Lower*	*Upper*
Intercept	*84.40*	*1.15*	*82.13*	*86.67*
Grade 11	*7.67*	*1.10*	*5.50*	*9.82*
Grade 12	*13.46*	*1.69*	*10.15*	*16.79*
σ_w^2 *(school)*	*13.81*	*6.14*	*3.43*	*26.27*
σ_u^2 *(student)*	*97.14*	*9.17*	*79.48*	*115.30*
σ_v^2 *(task)*	*10.89*	*3.96*	*4.37*	*18.84*
σ_ε^2	*198.92*	*6.90*	*185.58*	*212.60*

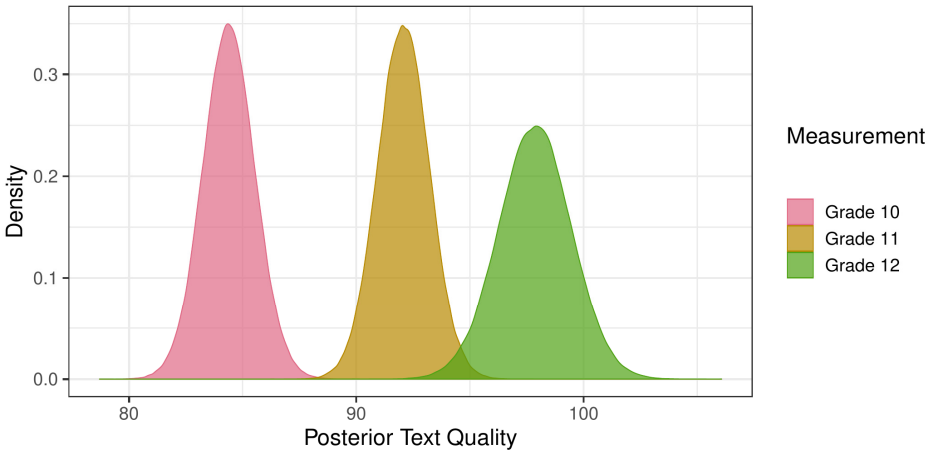


Figure 2. Posterior distribution of text Quality in grades 10, 11, and 12. Posterior distributions for grades 11 and 12 are obtained by adding the MCMC samples of the intercept to the MCMC samples for the improvement of the respective grade.

6. Experimental Data Set

Data were collected from 89 students of two high-schools in the Netherlands. Students made three writing tasks in one week; one on Monday, Wednesday, and Friday. Prior to writing the second and third text, the participants received feedback on their previously written text. As part of the feedback, students received annotated exemplar texts selected from the national baseline; these are texts that are representative of a specific position on the benchmark scale, accompanied with a description for several text quality aspects. Students could compare and contrast their own text with the exemplar texts of students with the same or a better score. Figure 3 depicts the data of the experiment per measurement occasion.

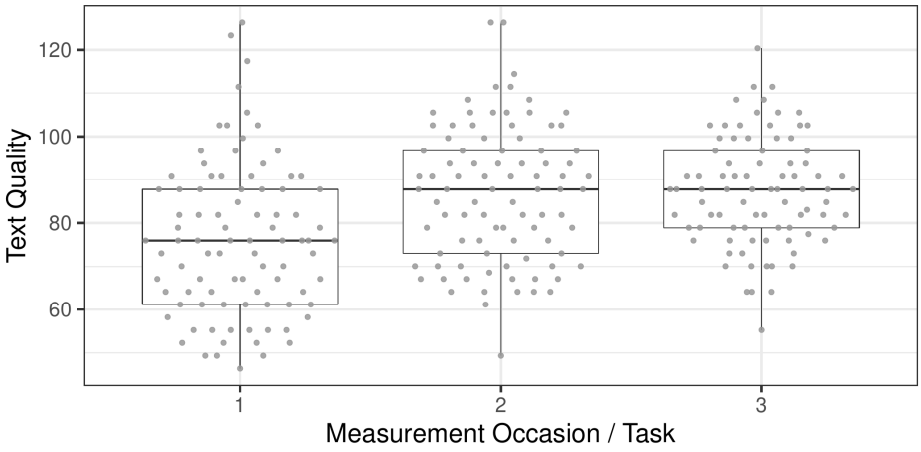


Figure 3. Box and whiskers plot of student performance on the three measurement occasions. Gray points represent the raw scores on text quality. Quasi-random jitter was added to the x-coordinates of the points to avoid visual clutter. The average performance clearly increases from measurement one to two, but it is hard to interpret this improvement without a reference group.

7. Experimental Analysis

A typical analysis for this data set is almost identical to that of the baseline data set, except that for this data set we estimate differences between measurements, which is contaminated with differences due to tasks. Therefore, we estimate fixed effects for measurement occasion and variance components for the differences between schools, students within school, and random error.[2] Since each student took only

[2] Let yhik be the observation of measurement h (h = 1, 2, 3) of student i (i = 1, ..., Ki) in school k (k = 1, 2). The multilevel model can be written as:

one task at each measurement occasion, the between-task variance cannot be estimated.

As for the baseline analysis, we summarize the posterior distribution of the multilevel model using the mean, standard deviation, and HPD in Table 2. This shows that the average text quality is estimated at 76.52 (95% HPD [64.61, 88.63]). At the second measurement occasion, students performed on average about 9.37 points better (95% HPD [5.59, 13.11]) than at intake. At follow up, students' improvement was estimated at 9.94 (95% HPD [6.03, 13.72]). A bivariate scatterplot for the parameters in Table 2 is shown in Figure A3 (appendix).

Table 2. Summary of the posterior distribution for the experimental data set. The first column shows the parameter, the second the posterior mean for that parameter, the third the posterior standard deviation and the last two columns show the 95% higher posterior density interval. The improvement of measurement 2 and 3 is relative to the intercept (measurement 1). The posterior standard deviation may be interpreted as a standard error.

			95% HPD	
Parameter	*Mean*	*SD*	Lower	Upper
Intercept	76.52	5.22	64.61	88.63
Measurement 2	9.37	1.92	5.59	13.11
Measurement 3	9.94	1.96	6.03	13.72
σ_w^2 (school)	86.86	213.12	1.77×10^{-10}	414.62
σ_u^2 (student)	101.10	23.68	57.24	148.63
σ_ε^2	144.50	15.91	115.00	176.40

$$y_{hik} = \beta_0 + \beta_1 \times [\text{Measurement}_{hik} = 2] + \beta_2 \times [\text{Measurement}_{hik} = 3] + w_{00k} + u_{0ik} + \epsilon_{hik}.$$

The fixed part consists of an intercept (β_0) that represents the mean writing score on the first measurement and two fixed effects that capture the difference in mean writing score between subsequent measurements and the first measurement (β_1 and β_2). The random part consists of a random intercept for school (w_{00k}), a random intercept for person within school u_{i0k} and a residual ϵ_{hik}.

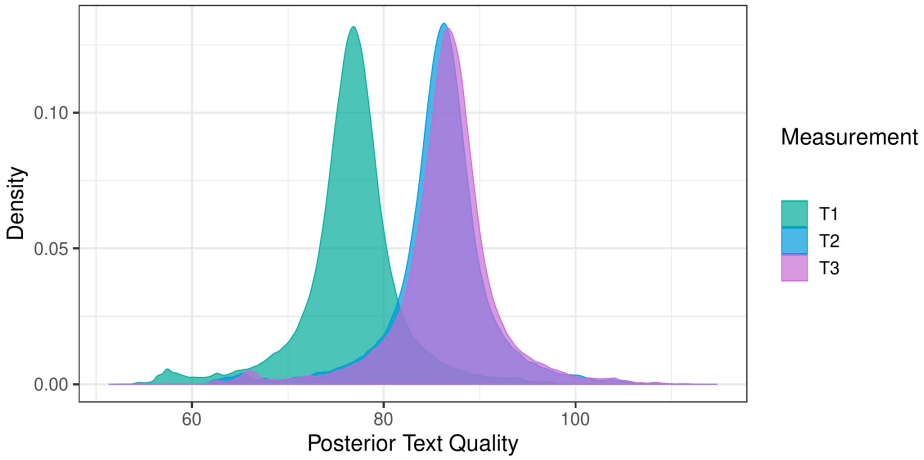


Figure 4. Posterior distribution of text quality at each measurement occasion of the product data set.

The estimated improvement across measurement occasions is shown in Figure 4. Apparent is that students perform better at posttest than at pretest and that the difference between follow-up and posttest seems almost negligible.

At this point in the analysis, drawing conclusions about the effect of the treatment is problematic because there is no control group. Thus, the improvement of the students cannot solely be attributed to just the intervention but might be caused by differences in difficulty between tasks.

8. Relating Baseline Results to the Analysis of an Experimental Study

Ideally, we directly compare the difference in text quality between measurement occasions in the experimental study. However, interpreting these differences is not straightforward as the contamination of task effect and measurement occasion makes this impossible. To make the differences between measurement occasions interpretable we need to correct these for task difficulty. As the baseline study provides estimates of task difficulty, a correction is self-evident. We can correct students' scores in the experimental study by subtracting the estimated task effect in the baseline study. As a consequence, the corrected posterior means for each measurement occasion changed slightly, see Figure 5 (from 76.52 to 79.53 for measurement 1, from 85.89 to 83.54 for measurement 2, and from 86.46 to 85.45 for measurement 3). Note that a direct comparison is possible because the rating procedure of the experimental study is identical to the rating procedure of the baseline study.

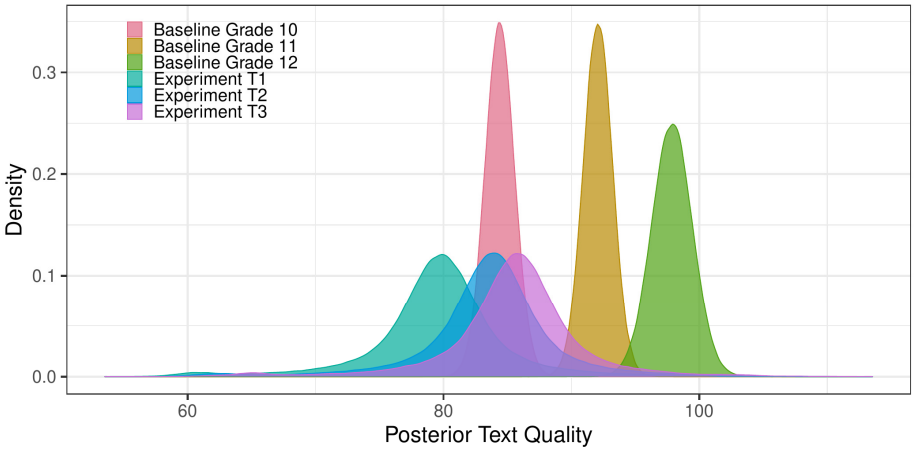


Figure 5. Posterior distributions for each grade in the baseline study and each measurement occasion in the experimental study. The posterior distributions of the baseline study are narrower because they are based on more observations. We subtracted the estimated average task effect of each task category in the baseline study from the posteriors distributions in the experimental study to correct these for the task effect.

From Figure 5 we can infer that the corrected difference between measurement 1 and measurement 2 in the experimental study is almost as large as the difference between grade 10 and 11 in the baseline study. Hence, there is a substantial difference between both measurements and thus an experimental effect. By comparison, the difference between measurement 2 and measurement 3 is much smaller. Of course, this is not a statistical test of significance. Typically, such a test should account for between-task variance. To obtain an estimate for the magnitude of between-task effects we can use the estimates of the baseline study to simulate a distribution of task difficulty. Next, we can compute the probability that the observed difference between measurement occasions in the experimental study is due to differences between tasks.

Since multilevel models typically assume that the random effects follow a normal distribution with mean 0 we simulate a large number of task effects from a normal distribution with mean 0. As variance for this normal distribution, we use the posterior samples for the between-task variance, to propagate the uncertainty in this parameter into the distribution over task-effects. Here, the first sampled task effects uses the first posterior sample of between-task variance, the second sampled task effect uses the second posterior sample of between-task variance, and

so forth.[3] In total 300,000 task-effects were sampled (the same amount as MCMC samples). Next, we can visualize the posterior distribution of improvement between measurement occasions and contrast this with the distribution over task-effects, as shown in Figure 6. The test presented here is similar to a z-test with a known population variance. Usually one computes the probability of the observed test statistic, a single value, assuming that the null hypothesis is true. The distribution of the test statistic under the null depends on the known population variance, the value to test against (e.g., 0), and the sample size. Here, the distribution under the null hypothesis is shown in grey. Rather than using a fixed value for the population variance we used the posterior distribution of task variance of the baseline analysis. This accounts for the uncertainty due to tasks (controlling for variance due to error, student, and schools). Likewise, rather than a single value for the test statistic, we use the posterior distribution of the effect of measurement occasion, which is represented by the colored distributions in Figure 6.[4]

The left and middle panel in Figure 6 contrast measurement occasions 2 and 3 against intake. The improvement appears to exceed what would be expected of a random task-effect. The right panel contrasts measurement 2 with measurement 3. Here, the improvement seems indistinguishable from a random task-effect.

The above results show that the observed effect between measurement occasions 1 and 2 is larger than can be expected from any given task. However, this does not provide a straightforward way in which to to interpret the magnitude of this effect.

---

[3] This sampling procedure approximates the following integral:

$$p(task\ effect \mid baseline) = \int p(task\ effect \mid task\ variance)\, p(task\ variance \mid baseline)\, d(task\ variance).$$

This is similar to simulating from the posterior predictive distribution of the baseline study.

[4] Alternatively, one could test this using Bayes factors. This requires approximating the posterior distribution of task effect in the baseline study by a parametric distribution and then using this as a prior distribution for the condition effect in the experimental study. Subsequently, an interval Bayes factor can be computed with the approach of Klugkist et al. (2005) using the 95% credible interval of posterior distribution of task effect. In this case, this yields the same conclusions as the z-test approach, see the supplementary materials at https://github.com/vandenman/Priors-Education.

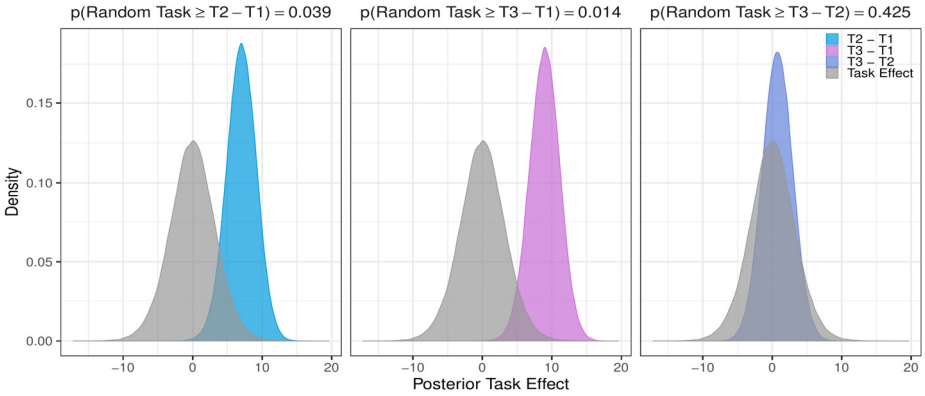


Figure 6. Distribution of the effect of a random task versus the posterior distribution of the estimated progress in the experimental data. The grey density represents the distribution of the effect of a random task. The blue density in the left panel is the posterior distribution of improvement between the first and second measurement; the purple density in the middle panel is the posterior distribution of improvement between the first and third measurement; the dark blue density in the right panel is the posterior distribution of improvement between the second and third measurement. The probability that a random task is larger than the improvement is shown above each panel.

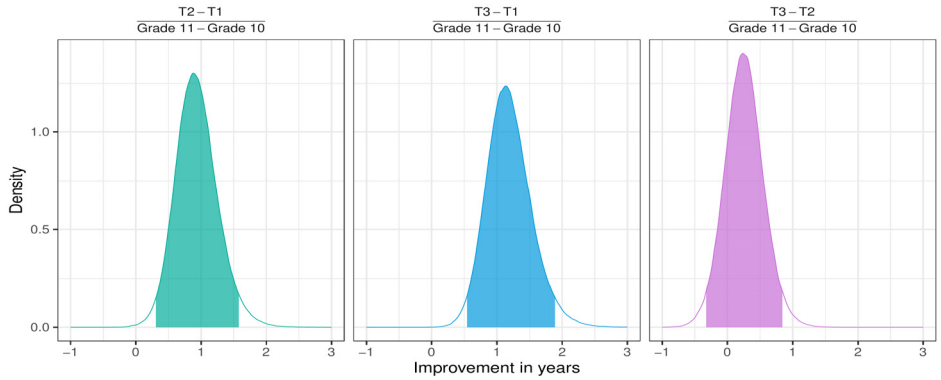


Figure 7. Improvement in the experimental study relative to the improvement between grades 11 and 10. The left panel shows the posterior distribution for the improvement between measurement 1 and measurement 2 divided by the improvement from grade 10 to 11 (95% HPD [0.31, 1.58]). The middle panel shows the improvement from measurement 1 to measurement 3, standardized in the same manner (95% HPD [0.54, 1.89]). The right panel shows the improvement between measurement 2 and 3 (95% HPD [–0.33, 0.84]). The shaded areas under the curve represent 95% HPD intervals.

To obtain a measure that is easy to interpret, we again compare the results to that of the baseline. From the baseline study, we obtained a posterior distribution that quantifies students' improvement between grade 10 and grade 11, accounting for differences between tasks (i.e., parameter Grade 11 in Table 1).

Next, we take the posterior samples for the effect of measurement in the experimental study (Parameter Measurement 2 in Table 2) and divide these by the samples of the baseline study. The resulting posterior distribution expresses the progress of students in the experimental study in baseline study years and provides a practically intuitive interpretation for the effect size. The resulting posterior distributions are shown in Figure 7.

In the left panel of Figure 7, the posterior has a mean 0.94 (95% HPD [0.31, 1.58]), which indicates that the students appear to have gained almost a year in ability between the two measurements. In the middle panel, the posterior mean is 1.19 (95% HPD [0.54, 1.89]) which indicates that this improvement is still present at the third measurement. In the far right-hand panel, the mean improvement is 0.25 (95% HPD [−0.33, 0.84]) which is much smaller and indicates less improvement between measurement occasions two and three.

9. Discussion

In this paper, we related the results of an experimental study to those of a baseline study. The post-intervention improvement in the experimental study exceeded the differences between random tasks in the baseline study. Therefore, we interpret the effect of the intervention as significant. We further quantified this effect by expressing the improvement in number of school years.

If a control group is missing, the task-effect cannot be disentangled from the effect of an intervention. However, by relating the increase in performance to estimates of the between-task variance in a baseline study, we can compute the probability that the improvement across measurements is a task-effect. This method could provide a point of reference for studies without a control group and may help discern between statistically significant effects and practically relevant effects and relate the effects of different studies to each other (Fan, 2001; Hojat & Xu, 2004).
A comparison with a baseline study can also enrich the results of studies with a control group. For example, if a study administers a smaller variety of tasks than a baseline study, a comparison can provide a clearer assessment of the generalization of an effect over tasks.

Comparing the distribution of task-effects in an experimental study to that of a baseline study relates to approaches of statistical tests for equivalence, such as TOST (Lakens, 2017), ROPE (Kruschke, 2011), and interval Bayes factors (Morey & Rouder, 2011). These three approaches have in common that a researcher specifies some minimal effect size below which an effect is practically equivalent to zero.

These tests also have in common is that they provide little guidance on how to determine such a minimal effect size. In contrast, our approach can be seen as deriving this minimal effect size from a baseline study (e.g., the effect size that is sufficiently implausible to be caused by between task effects).
Here, we opted to speak of significance when the posterior probability that the observed task-effect is larger than that of a random task is less than 0.05. This choice is arbitrary and other motivations have been suggested (Benjamin et al., 2018; McShane et al., 2019).

We opted for a Bayesian analysis because it allows us to account for the uncertainty in the estimates of the baseline study when comparing these to the results in the experimental study. To some extent, it is also possible to do this in a frequentist analysis. For example, a distribution of effect sizes can be approximated using a normal distribution with as mean the point estimate and as standard deviation the corresponding standard error.

Typically, comparing a baseline with experimental data is an ill-advised procedure because of the absence of sampling equivalence of groups (Campbell, 1957, p. 300), differences in incentives (Braun et al., 2011), or the use of different measurement instruments (Van den Bergh et al., 2012). Relating different data sets always requires argumentation of comparability, that is, both the population and the instruments must be comparable. In the present study, baseline and experimental data come from the same population of students (e.g., same grade and same school track). Moreover, the writing prompts in the experimental study were a subset of those in the baseline study. Therefore, we assume that the baseline and experimental are sampling equivalent.

10. Limitations

As is typical for quasi-experimental research in educational settings, there was no random assignment of students to conditions in the experimental study. Therefore, the usual limitations of quasi-experimental research apply; it is possible that the observed differences between measurement occasions are caused by a confounding variable rather than the intervention because students are not randomly assigned to either the experimental or baseline condition/ study. Caution should be exercised in interpreting the conclusions based on a comparison with a baseline study as causal.

A key assumption of multi-level models is that task-effects are, at least asymptotically, normally distributed. If normality is violated then the probabilities shown in Figure 6 could be biased. Here, we briefly outline an argument on why the task effects are likely approximately normally distributed. Note that a naive estimator for the effect of a task is simply the mean of the students' scores on that task. Although this estimator is unbiased, much better estimates can be obtained by accounting for the hierarchical structure (e.g., Efron & Morris, 1977). Since the naive

estimator is an average the central limit theorem applies and thus the distribution of task-effects converges asymptotically to a normal distribution (under mild regularity conditions).

Another avenue for incorporating the results of a baseline study into the analysis of an experimental study is through the prior distribution. The posterior distribution of the baseline study could serve as the prior distribution for the experimental study.

Although this is conceptually straightforward, we did not do so for two reasons. First, to obtain exact approximations to the posterior, the analyst of the experimental study must have the original data to obtain posterior distributions for the baseline data set. In practice, it is unlikely that an analyst has access to a baseline data set which limits the applicability of the method. It is possible to approximate the marginal posterior distributions using some parametric family of distributions, which can then be published and used in experimental studies. However, these approximations will likely ignore the correlations and other higher-order moments in the posterior distribution. The consequences of ignoring the higher-order moments in the posterior distribution are simply unknown. Second, the benefit of informed priors is unclear, as the data typically overwhelm the influence of the prior distribution, barring extreme cases (Lynch, 2007). Thus, since the inferences drawn in the paper are based solely on the posterior distribution, the influence of the prior distribution is likely negligible.

11. Recommendations

Prior information can enrich statistical analyses and provide more insight into the data. Here, we outline three recommendations for those who wish to apply our method for incorporating prior information in practice.

A key requirement for comparing results from a large-scale assessment with those of an experimental study is that the data are comparable in terms of populations and tests. Especially the comparability of writing research hinges on the validity of the measurement instruments (Graham & Harris, 2014). That the instruments measure what they are supposed to measure (the traditional definition of validity) is not as important as that they measure the same construct. If the baseline study measured different constructs than the experimental study, for instance, because different measurement instruments were used, then a comparison is unintelligible and thus meaningless. Thus, to make a meaningful comparison with baseline results we recommend using the same measurement instruments as those used in a baseline assessment.

The use of a baseline study instead of a control group opens up new avenues for designing experimental studies. Currently, researchers tend to allocate about half of the available resources to a control group and the other half to an experimental group. However, since the experimental group can now be related to

a baseline study, it becomes possible to allocate funds to a theoretically competing theory, rather than to a control group. We recommend considering the possible use of national assessment data when designing a study.

The use of our method is depends on large-scale assessments publishing their results. It is key that those studies either disclose the raw data or publish the marginal posterior distributions of the parameters. If the results of large-scale assessments are not available as a benchmark, then it is inoperable to use them to inform the analysis of experimental studies. Therefore, we recommend making data available, either as raw data or in a summarized form.

In sum, we related the results from a baseline study to the analysis of an experimental study that lacked a control group. This allowed us to determine whether the differences between measurements in the experimental group exceeded what would be expected from between task variance. Altogether, this may help to place effect sizes of experimental studies in a broader context.

Acknowledgements

Code for the analyses and a simulated data set is availabe on GitHub at https://github.com/vandenman/Priors-Education

References

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-.-J.,Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D.,Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A.,Easwaran, K., Efferson, C., . . . Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6–10. https://doi.org/10.1038/s41562-017-0189-z

Blok, H. (1986). Essay rating by the comparison method. *Tijdschrift voor onderwijsresearch*, *11* (4), 169–176.

Bouwer, R., Koster, K., & van den Bergh, H. (2017). Leren schrijven met teksters: Een wetenschappelijk beproefde lesmethode voor het basisonderwijs [Learning to write with tekster: A scientifically proven teaching method for elementary schools]. *Pedagogische studiën*, *94* (4), 304–329.

Bouwer, R., Béguin, A., Sanders, T., & Van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing*, *32* (1), 83–100. https://doi.org/10.1177/0265532214542994

Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade naep reading assessment. *Teachers college record*, *113* (11), 2309–2344. https://doi.org/10.1177/016146811111301101

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80* (1), 1–28. https://doi.org/10.18637/jss.v080.i01

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological bulletin*, *54* (4), 297. https://doi.org/10.1037/h0040950

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76* (1). https://doi.org/10.18637/jss.v076.i01

De Smedt, F., Van Keer, H., & Merchie, E. (2016). Student, teacher and class-level correlates of flemish late elementary school childrens writing performance. *Reading and writing*, *29* (5), 833–868. https://doi.org/10.1007/s11145-015-9590-z

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, *236*, 119–127.

Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. *The Journal of Educational Research*, *94* (5), 275–282. https://doi.org/10.1080/00220670109598763

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, *1* (3), 515–534.

Graham, S. E., & Harris, K. R. (2014). Conducting high quality writing intervention research: Twelve recommendations. *Journal of Writing Research*, *6* (2), 89–123. https://doi.org/10.17239/jowr-2014.06.02.1

Hojat, M., & Xu, G. (2004). A visitor's guide to effect sizes–statistical significance versus practical (clinical) importance of research findings. *Advances in health sciences education*, *9* (3), 241–249. https://doi.org/10.1023/B:AHSE.0000038173.00909.f6

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59* (1), 57–69. https://doi.org/10.1111/j.1467-9574.2005.00279.x

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6* (3), 299–312. https://doi.org/10.1177/1745691611406925

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, *8* (4), 355–362. https://doi.org/10.1177/1948550617697177

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. Springer.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*, 235–245. https://doi.org/10.1080/00031305.2018.1527253

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*, 406–419. https://doi.org/10.1037/a0024377

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rietdijk, S., Janssen, T., van Weijen, D., van den Bergh, H., & Rijlaarsdam, G. (2017). Improving writing in primary schools through a comprehensive writing program. *The Journal of Writing Research*, *9* (2), 173–225. https://doi.org/10.17239/jowr-2017.09.02.04

Rijlaarsdam, G., Van den Bergh, H., & Zwarts, M. (1992). Incidentele transfer bij produktieve taalopdrachten: Een aanzet tot een baseline [Incidental transfer on productive language tasks: An initiation for a baseline.] *Tijdschrift voor Onderwijsresearch*, *17*, 55–66.

Rijlaarsdam, G., Van den Bergh, H., Couzijn, M., Janssen, T., Braaksma, M., Tillema, M., Graham, S., Bus, A., Major, S., & Swanson, L. (2012). Writing. In K. R. Harris, S. E. Graham, T. E. Urdan, A. G. Bus, S. E. Major, & H. Swanson (Eds.), *APA educational psychology handbook, Vol 3: Application to learning and teaching*. (pp. 189–227). American Psychological Association. https://doi.org/https://doi.org/10.1037/13275-000

Van den Bergh, H., De Maeyer, S., Van Weijen, D., & Tillema, M. (2012). Generalizability of text quality scores. *Measuring writing: Recent insights into theory, methodology and practices*, *27*, 23–32. https://doi.org/10.1163/9789004248489_003

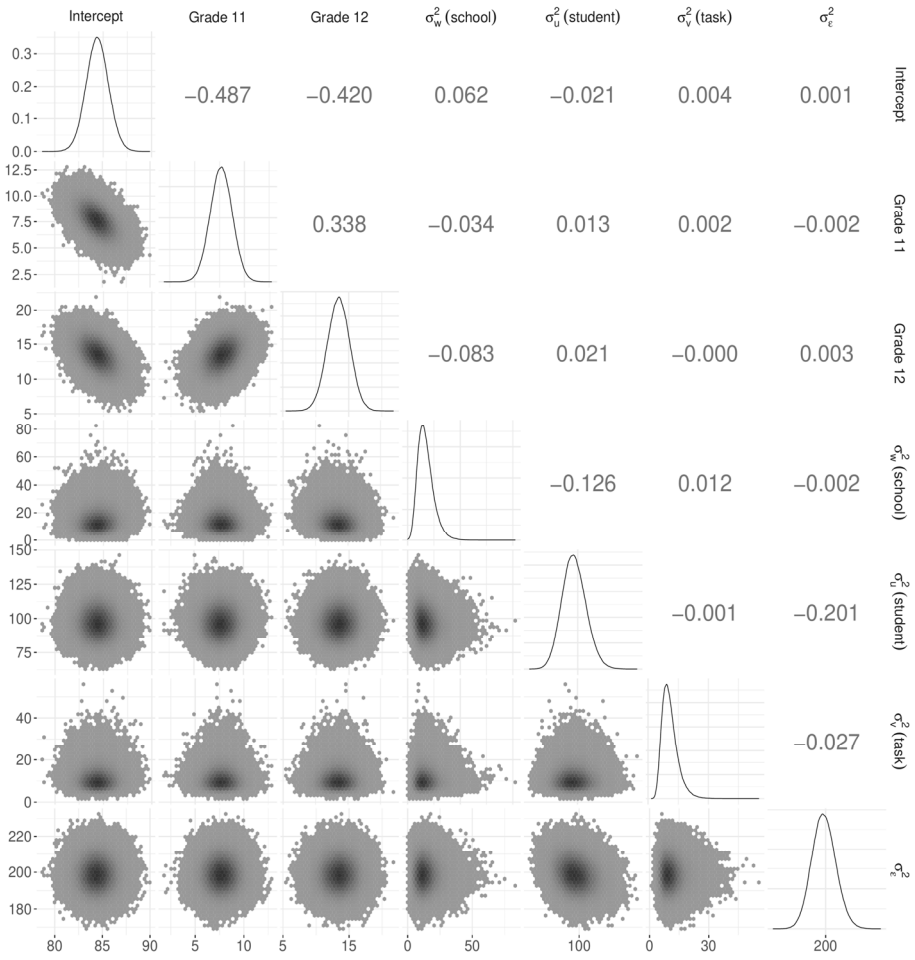
Van den Bergh, H., & Eiting, M. H. (1989). A method of estimating rater reliability.
Journal of Educational Measurement, *26* (1), 29–40. https://doi.org/10.1111/j.1745-3984.1989.tb00316.x

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (Eds.). (2018). Editorial: Bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review*, *25*, 1–4. https://doi.org/10.3758/s13423-018-1443-8

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P. C. (2021). Rank-normalization, folding, and localization: an improved R for assessing convergence of MCMC (with discussion). Bayesian analysis, 16(2), 667-718. https://doi.org/10.1214/20-ba1221

Zwarts, M., Rijlaarsdam, G., Janssens, F., Wolfhagen, I., Veldhuijzen, N., & Wesdorp, H. (1990). Balans van het taalonderwijs aan het einde van de basisschool [Balance of language teaching at the end of the elementary school]. *Uitkomsten van de eerste taalpeiling einde basisonderwijs*. https://doi.org/10.1163/2214-8264_dutchpamphlets-kb2-kb29970

Appendix A: Convergence Diagnostics and Visual Summaries



Figureurerabic A1. Visual summary of the posterior distributions for the group Level effects of the baseline data set. The strips above and right of the figures indicate the parameters compared. Figures on the diagonal show marginal density estimates. Figures below the diagonal show bivariate hexagonal histograms. The numbers above the diagonal indicate the Pearson correlation between the samples of the parameters.

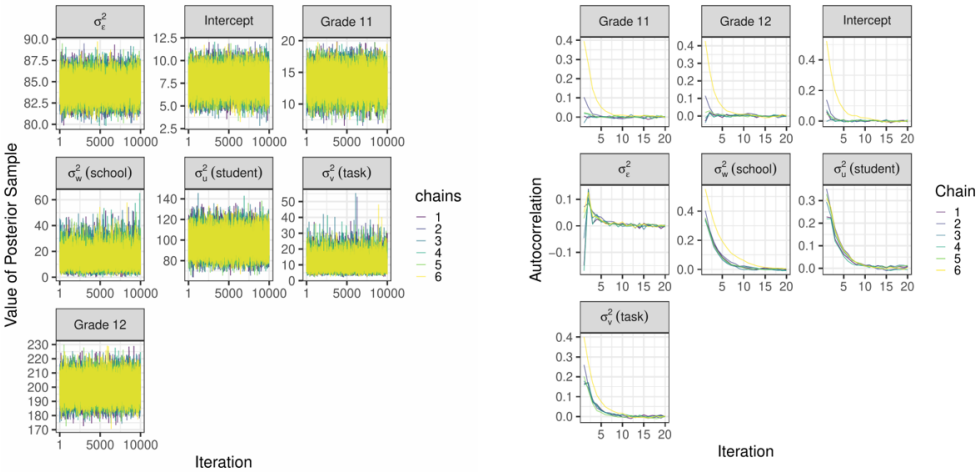


Figure 2. Convergence diagnostics for the analysis of the baseline data set. Left: trace plots of the first 10,000 posterior samples after warmup. The different chains appear indistinguishable, which indicates they converged. Right: Autocorrelation of the chains. The 0th lag was omitted (as this is 1 by definition). The autocorrelation drops to 0 after about 5 iterations.

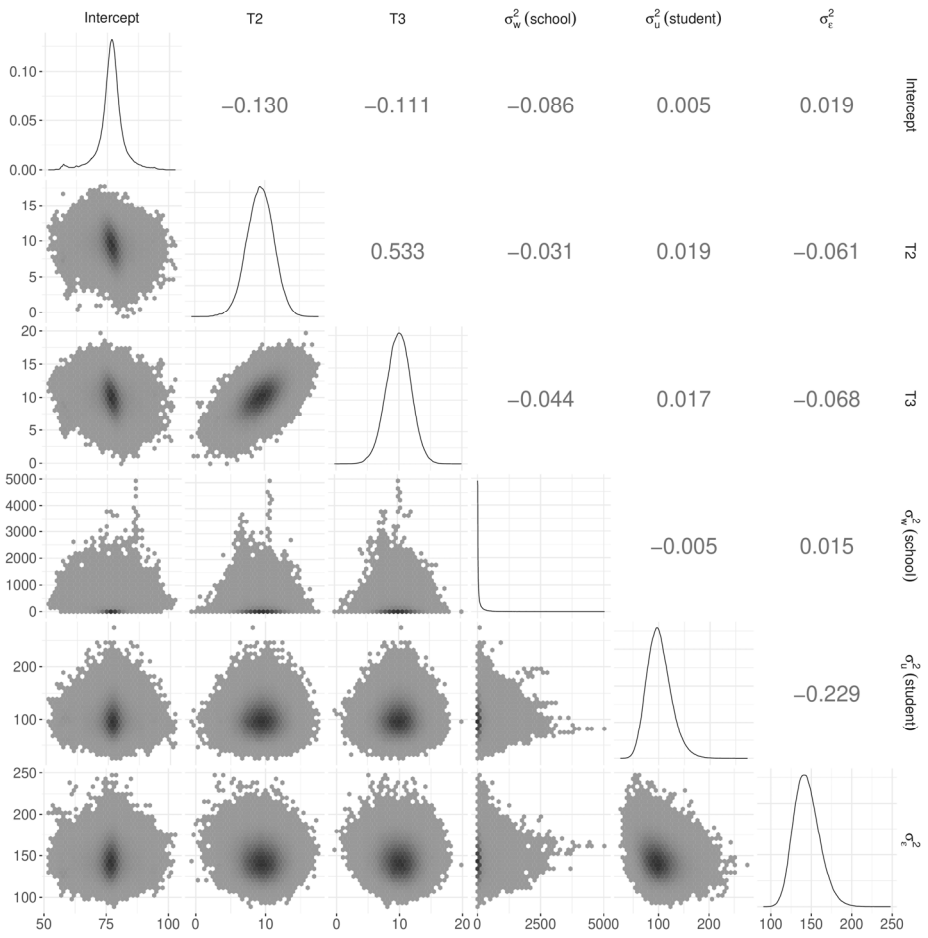


Figure 3. Visual summary of the posterior distributions for the group level effects of the experimental data set. The strips above and right of the figures indicate the parameters compared. Figures on the diagonal show marginal density estimates. Figures below the diagonal show bivariate hexagonal histograms. The numbers above the diagonal indicate the Pearson correlation between the samples of the parameters.

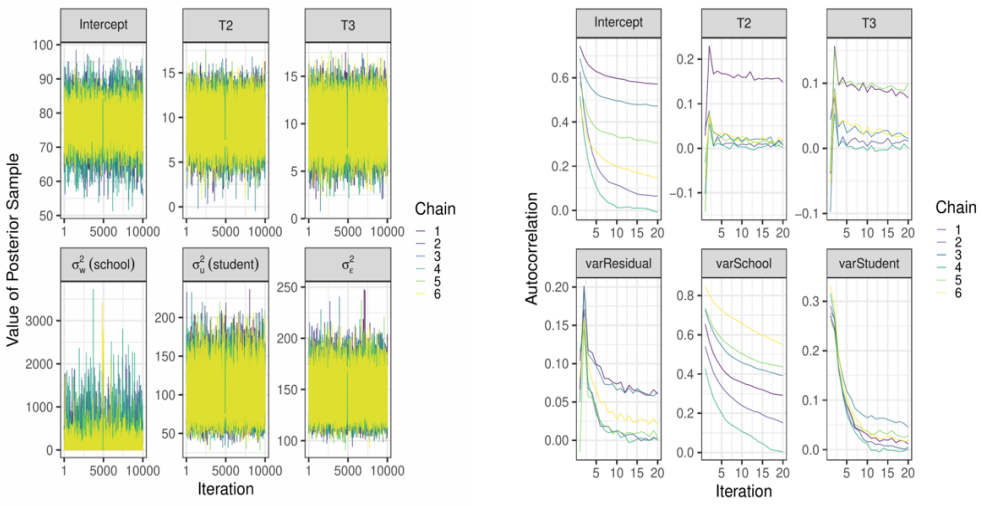


Figure A4. Convergence diagnostics for the analysis of the experimental data set. Left: trace plots of the first 10,000 posterior samples after warmup. The different chains appear indistinguishable, which indicates they converged. Right: Autocorrelation of the chains. The 0th lag was omitted (as this is 1 by definition). The autocorrelation drops to 0 after about 10 iterations.