Todo list

Priors Information for Multilevel models in Educational Analyses

Don van den Bergh*[1], Nina Vandermeulen[2], Rianne[2], Marije Lesterhuis[2], Sven de Maeyer[2], Elke van Steendam[2], Gert Rijlaarsdam[2], and Huub van den Bergh[3]

[1]University of Amsterdam
[2]University of Antwerp
[3]University of Utrecht

Abstract

Scholastic achievement is often monitored in national assessments. For each topic there are multiple tasks that measure a student's skill, to avoid task-specific effects. For example, reading comprehension is measured with multiple texts. In experimental studies, a similar approach is often employed; because student performance varies across tasks, multiple tasks are administered so that the influence of a experimental condition may be assessed while controlling for the effect of task. National assessments consist of many tasks and the data is generally rich in information. In contrast, experimental studies typically use only one assignment, which makes it difficult to distinguish improvement in an experimental condition from between-task variance. This prompts the question whether knowledge obtained from national assessments about between-task variance can be used in the analyses of experimental studies. Here, we demonstrate how information of a baseline data set can be used in the analysis of an experimental study. We adopt a Bayesian paradigm as this enables us to propagate the uncertainty in the estimates of a national assessment into the analysis of the experimental study.

*Correspondence concerning this article should be addressed to: Don van den Bergh, University of Amsterdam, Department of Psychological Methods, Postbus 15906, 1001 NK Amsterdam, The Netherlands. E-Mail should be sent to: donvdbergh@hotmail.com.

Don: 0. Specifeer propere priors. 1. Herhaal alle analyses met brms. 2. Transformeer alle standaard deviaties naar varianties direct na het samplen en sla dat op. 3. Schoon de code op wat betreft het transformeren van varianties.

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). For instance, in the Netherlands every four years the achievements of students are measured at the end of primary education in some of the most important subject areas. Although the results of these assessments often inform policymaking, the data are seldom used in educational research even though there are ample opportunities.

A common denominator in national assessments is for all subject areas measurements are based on an analysis of that subject area. Therefore, students read multiple texts if reading is assessed or write multiple texts if writing is assessed. This is a necessity if one wants to describe the level of achievements covering a whole domain while generalizing over specific assignments (or tests) at the same time. For writing in the Netherlands, for instance, students wrote 21 different texts in a national assessment (Zwarts et al., 1990). Of course, not all students take every test, but a sparse design is in operation, in order to minimize testing time but allowing for conclusion at 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 pretest and one text as posttest (e.g., Graham & Harris, 2014). Based on these texts we are proned to draw conclusions on changes in the writing skills of students. Although it is well documented that differences between different types of writing assignments can be large (e.g., Bouwer, Béguin, Sanders, & Van den Bergh, 2015), and we hardly can make inferences based on only one writing assignment. Of course, many researchers are aware of the limited generalizability of single-task experiments. However, it is often infeasible that students write more tasks.

So, on the one hand there is much information on levels of achievement of students (at certain levels of education) from assessments, and on the other hand in many experimental studies we rely on relative small samples and relative narrow measures of skills. Therefore, one could wonder why don't we use the information from large scale assessments? Can this information from assessments be put to use in our experimental studies?

In fact, the information from national assessments can be seen as information on the level of achievements in general. In this sense, this information might be seen as prior knowledge that describes the standard level of achievements. In experimental studies we like to show that the increase in achievements due to the experimental manipulation exceeds 'natural' growth. Therefore, the information from assessments might function as a baseline, or standard level of achievement for experimental studies. Second, results of prior studies form the basis for new studies and research hypotheses. Nevertheless, we do not fully use the available data. Prior knowledge, prior data, is rarely used in statistical analyses. This

might be inefficient, as we keep measuring students over and over again in order to get studies which have enough power to draw conclusions. However, we can also increase the power of studies if we enrich our analyses with prior results (Graham & Harris, 2014). One of the type of studies that comes directly to mind are of course assessment studies, not only because many students take these tests, but also because students take many tests in order the generalize over the idiosyncrasies of specific tests.

Unfortunately, there exists no straightforward method to incorporate prior information into (frequentist) 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 this values. However, this approach seems far from ideal, as measurement error and uncertainty in the prior results are completely ignored. Although such uncertainties could be introduced by means of standard errors, many types of frequentist analyses are not equipped for such procedures.

In this respect, Bayesian analyses of prior results from the data might be preferable; Bayesian statistics offer a rigorous and consistent approach to quantify uncertainty in statistical analyses. In Bayesian inference, prior knowledge (or a lack thereof) is represented by probability distributions, which describe all uncertainty about the quantities of interest. Upon observing the data, prior knowledge is updated to posterior knowledge, which is again represented by probability distributions. Key is that these probability distributions provide a complete account of the uncertainty. Thus, Bayesian inference is an ideal vehicle to reuse findings from prior analyses into future studies, while accounting for the uncertainty in these prior results. In educational research, there is an abundance of data, but results from the analyses are rarely used in the analysis of new studies.

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. By means of this baseline data set, we provide a brief explanation of Bayesian statistics, before analyzing the data with a multilevel model. Next, we analyze a follow up data set and relate the parameter estimates from the baseline analysis to those of the follow up. The paper is concluded with a discussion on the widespread applicability and benefits of this approach and the limitations concerning the validity of this approach.

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, 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 students made four at random. Some students did not make all tasks, 497 students made four tasks and 128 students made three or fewer tasks. The minimum amount of students per task was 62 whereas the maximum was 84. Students' text were rated by... .

Don: Aanvullen aub

The data contain an obvious nested structure, as illustrated in Figure 1. Observations are nested within students and tasks. Furthermore, students are nested within schools. Students took a random sample of four tasks out of 32 tasks developed for this writing assessment.

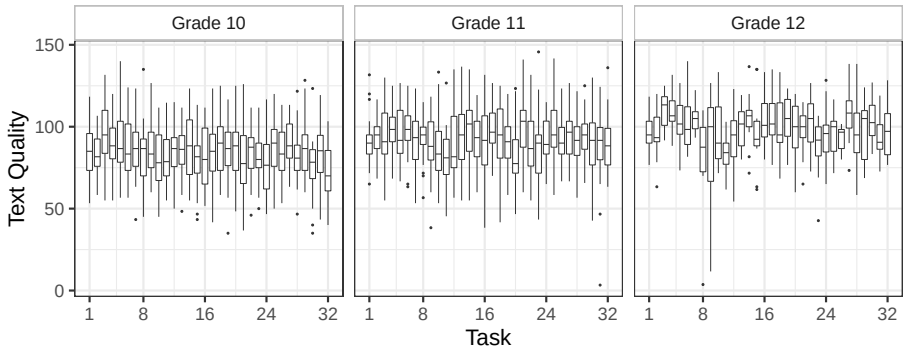

Figure 1: Box and whiskers plot of student performance on each task for the three grades measured. The grade is indicated above each panel and the task code is shown on the x-axis. There is substantial variance in student performance between tasks and within tasks, and student performance appears to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

The observations of text quality cannot be considered as independent. 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 score on different writing tasks. Therefore, a cross classified multilevel model is in operation. 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 to be analyzed as:

$$y_{(ij)k} = \beta \times \text{Grade}_{ijk} + [w_{00k} + u_{i0k} + v_{0j0} + \epsilon_{(ij)k}].$$

The model consists of two parts: a fixed parts and a random part (between square brackets). In the fixed part, Grade_{ijk} is an indicator matrix for students' grade. Consequently the vector of regression weights (β) represent the mean writing score for each grade. In the random part four residual scores are distinguished, all of which are assumed to be normally distributed around with an expected value of 0. The first residual (w_{00k}) captures the difference between a school and the average. The second residual (u_{i0k}) captures that the average score of student i in school k can deviate from the schools' mean. The third residual (v_{0j0}) captures that some tasks might be more difficult than other 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.

Bayesian Inference

This section aims to give a brief introduction to Bayesian inference with an emphasis on the problem at hand. For a more elaborate introduction to Bayesian inference, see the recent special issue in *Psychonomic Bulletin & Review* which provides tutorials and guidance for aspiring Bayesians (Vandekerckhove, Rouder, & Kruschke, 2018). The choice for a Bayesian analysis is motivated by the fact that Bayesian inference is naturally accompanied by uncertainty estimates, as is explained later. Thus, the estimates of a baseline study and an experimental study can be compared while accounting for the uncertainty in both sets of estimates.

Bayesian inference is centered on the updating of beliefs. For any parameter in a given statistical model $\mathcal{M}$, the values this parameter can take are assigned a prior belief. These beliefs are represented with a probability distribution, usually called the prior distribution π. For example, in a multilevel model, the intercept β_0 can be assigned a normal distribution as prior distribution with mean 0 and variance 1. Then the a-priori the most likely values for the intercept are near 0 and about 95% of the prior mass lies within -1.96 and 1.96.

The key step in Bayesian inference is to use the data $\mathcal{D}$ to update the prior beliefs to posterior beliefs. The procedure for updating the prior distribution to a posterior distribution is given by Bayes theorem:

$$\underbrace{p\left(\boldsymbol{\beta} \mid \mathcal{D}, \mathcal{M}\right)}_{\text{Posterior}} = \overbrace{\pi\left(\boldsymbol{\beta} \mid \mathcal{M}\right)}^{\text{Prior}} \underbrace{\overbrace{\frac{p\left(\mathcal{D} \mid \boldsymbol{\beta}, \mathcal{M}\right)}{p\left(\mathcal{D} \mid \mathcal{M}\right)}}^{\text{Likelihood}}}_{\substack{\text{Marginal} \\ \text{Likelihood}}}.$$

Here, $\boldsymbol{\beta}$ represents all parameters in the model. The prior distribution of the parameters is updated through the likelihood of the statistical model. The likelihood is divided by the marginal likelihood so that the posterior distribution is a proper probability distribution (i.e., it integrates to 1). The posterior distribution is key for parameter estimates. For instance, if a single estimate for a parameter is desired, one could use the mean of the posterior distribution. Other often-used point-estimates are the posterior mode and posterior median. Simultaneously with obtaining the posterior, a measure of uncertainty for each parameter is obtained. Since the posterior distribution is a proper probability distribution, we can make inferences about the parameters. For example, given the posterior distribution for the intercept, $p\left(\beta_0 \mid \mathcal{D}, \mathcal{M}\right)$. This implies questions such as "Given that we have seen the data, what is the probability that the intercept is larger than 0?" Can be answered by computing $p\left(\beta_0 > 0 \mid \mathcal{D}, \mathcal{M}\right)$. Likewise, if we find a lower bound LB and upper bound UB for the intercept β_0 such that $p\left(LB \leq \beta_0 \leq UB \mid \mathcal{D}, \mathcal{M}\right) = 0.95$, we can claim: "Given that we have seen the data, we are 95% confident that the true value of the intercept lies between LB and UB." This interval is known as the Bayesian 95% credible interval. Another often-used Bayesian uncertainty interval is the 95% highest posterior density interval (HPD), an interval that contains 95% of the posterior mass and has the values of highest probability density.

Approximations to Posterior Distributions

Although Bayes theorem may appear straightforward, in practice the posterior distribution can be a high-dimensional probability distribution that is difficult to study analytically. Rather than studying the mathematical form of the posterior, it is much easier to simulate random values from the posterior distribution and to use these for inference. Such simulation methods are commonly referred to as Markov chain Monte Carlo (MCMC). The idea is that instead of computing a statistic of the posterior distribution in closed form, we can draw random observations from the posterior and use a sample estimator to approximate the statistic of the distribution. For example, if we are interested in the posterior mean of the intercept, we simulate many observations from the posterior distribution and use the sample mean of these observations to approximate the posterior mean of the intercept. Likewise, to compute the posterior probability that an intercept β_0 is positive, $p(\beta_0 > 0 \mid \mathcal{D}, \mathcal{M})$, we examine the proportion of MCMC samples where β_0 is positive. This procedure is akin to how applied scientists attempt to randomly sample participants from a population and then generalize the sample statistics to the entire population, with the exception that is it relatively easy to draw enormous samples with MCMC to obtain near-perfect approximations.

Statistical Software

All analyses were done in R (R Core Team, 2019). 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). For all analyses, we used six MCMC chains to assess convergence. 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.

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. However, the estimated variance between schools (13.81), students within school (97.14), and tasks (10.89) clearly deviate from 0. Since the data set contained such a large variety of schools and tasks, it is likely that these findings 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.

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).

			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

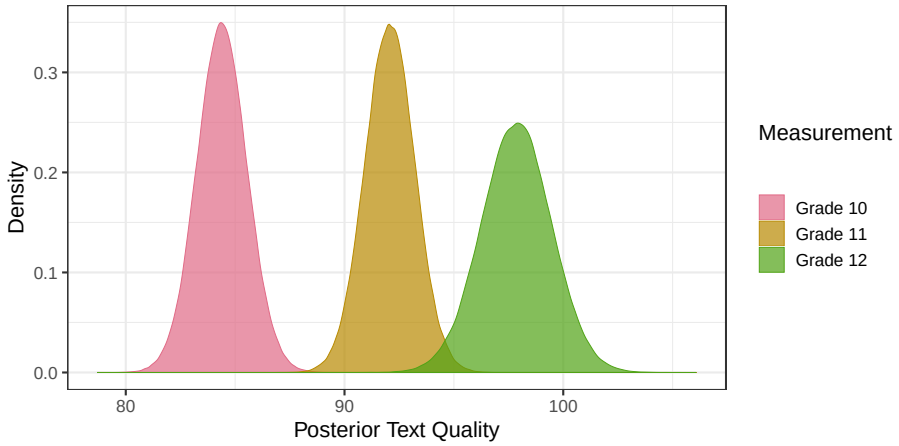

Figure 2: Posterior distribution of text quality in grades 10, 11, and 12. Posteriors distributions for grades 11 and 12 are obtained by adding the MCMC samples of the intercept to the MCMC samples for the improvement of the respective grade.

Application to an Experimental Analysis

Data Set

Data was 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. After the first task, the students received feedback on the quality of their written texts by a rating scale based on the baseline data set.

Don: Aanvullen aub!

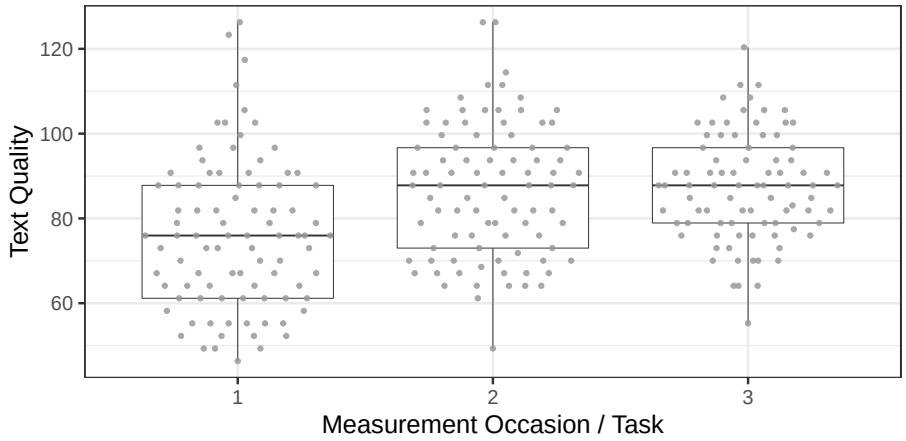

Figure 3: Box and whiskers plot of student performance on the three measurement occasions. Grey points represent the raw scores on text quality. Quasi-random jitter was added to the x-coordinates of the points to avoid visual clutter. The average performance clearly increases from measurement one to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

Analysis

A typical analysis for this data set is almost identical to that of the baseline data set, except that here we estimate differences between measurements, which might be contaminated with differences due to tasks. Since each student took only one task at each measurement occasions, the between-task variance cannot be estimated. Thus $y_{(hi)k}$ is the observation of measurement h ($h = 1, 2, 3$) of student i ($i = 1, ..., K_i$) in school k ($k = 1, 2$). The multilevel model thus becomes:

$$y_{(ij)k} = \beta_0 + \beta \times \text{Measurement}_{ijk} + [w_{00k} + u_{i0k} + \epsilon_{ijk}].$$

Here, y_{ijk} is the observation of student i on measurement j in school k. The fixed part consists of an intercept (β_0), fixed effect of measurement β. 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_{ijk}. 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 75.90 (95% HPD [65.62, 87.71]). At the second measurement occasion, students performed on average about 10.40 points better (95% HPD [6.81, 13.93]) than at intake. At follow up, students' improvement was estimated at 11.00 (95% HPD [7.48, 14.62]). A bivariate scatterplot for the parameters in Table 2 is shown in Figure 10.

The estimated improvement across measurement occasions is shown in Figure 4. Apparent is that students perform better at post-test than at pre-test and that the difference between follow up and post-test appears 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

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).

			95% HPD	
Parameter	Mean	SD	Lower	Upper
Intercept	75.90	4.72	65.62	87.71
Measurement 2	10.40	1.82	6.81	13.93
Measurement 3	11.00	1.82	7.48	14.62
σ_w^2 (school)	95.26	249.39	$9.78 \cdot 10^{-9}$	472.30
σ_u^2 (student)	101.19	23.65	57.11	148.11
σ_ϵ^2	144.11	15.71	114.34	175.31

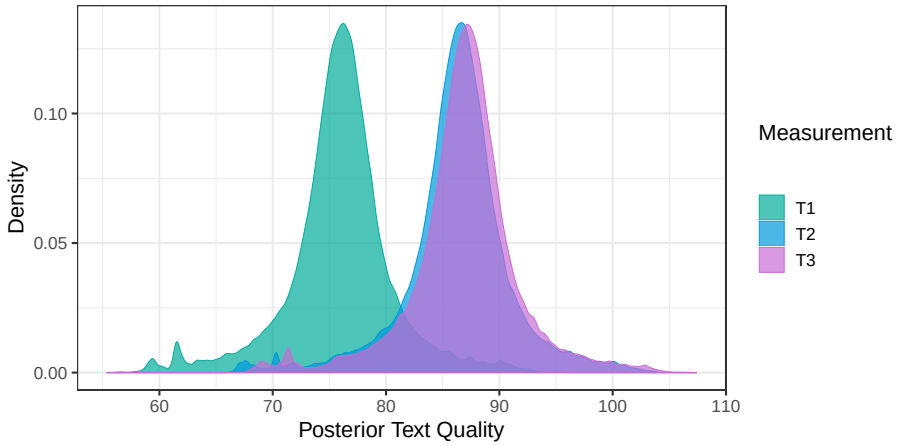

Figure 4: Posterior distribution of text quality at each measurement occasion of the product data set.

caused by differences in difficulty between tasks.

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 make this impossible. To make the differences between measurement occasions interpretable we need to correct these for task difficulty. As the 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 in the baseline study. As a consequence, the corrected posterior means for each measurement occasion changed slightly, see Figure 5 (from 75.90

to 78.91 for measurement 1, from 86.30 to 83.95 for measurement 2, and from 86.90 to 85.88 for measurement 3). Note that a direct comparison is possible because the rating procedure of the experimental study is based on the rating procedure of the baseline study.

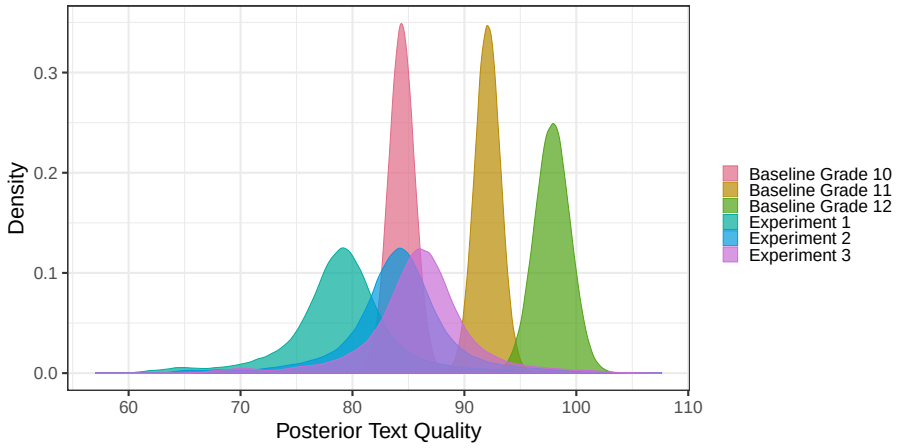

Figure 5: Posterior distributions for each grade in the baseline study and each measurement occasion in the experimental study. The posterior distributions of the baseline study are more narrow because they are based on more observations. We subtracted the estimated average task effect of each task category in the baseline study from the posteriors distributions in the experimental study to correct these for 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 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 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 these from a normal distribution. As variance, we use the posterior samples for the between-task variance, to propagate the uncertainty in this parameter into the distribution over task-effects.[1] In total 300,000 task-effects were samples (the same amount as MCMC samples). Next, we can visualize the posterior distribution of improvement between measurement occasions and contrast this to the distribution over task-effects, as shown in Figure 6.

Don: tot hier gekomen!

The left and middle panel in Figure 6 contrast measurement occasions 2 and 3 against intake. The improvement appears to exceeds what would be ex-

Don: Dit moeten we ook voor de prior verdelingen berekenen, en vervolgens kiezen we de prior zo dat deze kans a-priori 0.5 is. Vervolgens kunnen we aan de hand van de prior en posterior kans een BF berekenen. Op dit moment is de prior op beta improper, maar als we daar een Cauchy(0, 1)

[1] Essentially, the effects of these new random tasks are drawn from the posterior predictive distribution of the baseline study.

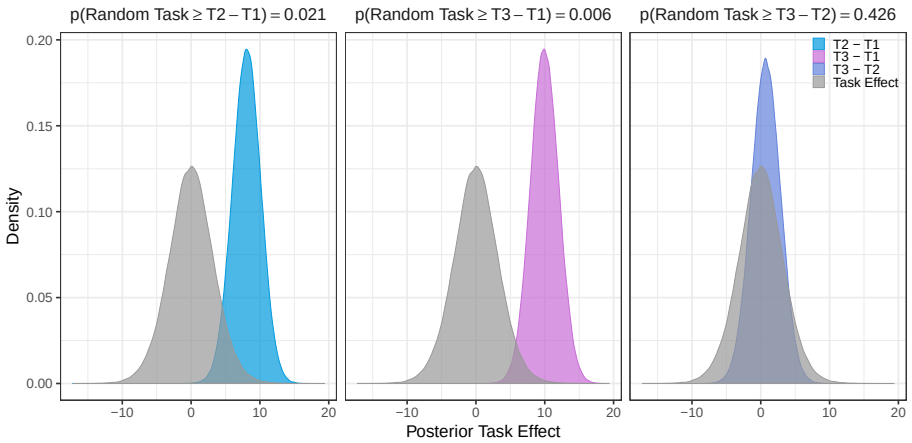

Figure 6: Distribution of the effect of a random task (grey) versus the posterior distribution of the estimated progress in the experimental data. The blue density in the left panel is the the posterior distribution of improvement between the first and second measurement; the purple density in the middle 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 large than the improvement is shown above each panel.

pected of a random task-effect. The right panel contrasts measurement 2 with measurement 3. Here, the improvement seems indistinguishable from random task-effect. This makes sense as there was no intervention between measurements 2 and 3.

The above results show that the observed effect between measurement moment 1 and 2 is greater than can be expected from any given task. However, this does not provide a clear way to straightforward manner to interpret the magnitude of this effect. To obtain an 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 experimental study in baseline study years and provides a practically intuitive interpretation for the effect size.

Discussion

In this paper, we introduced a procedure for comparing results from a large scale assessment into the analysis of an experimental study that lacked a control group. 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

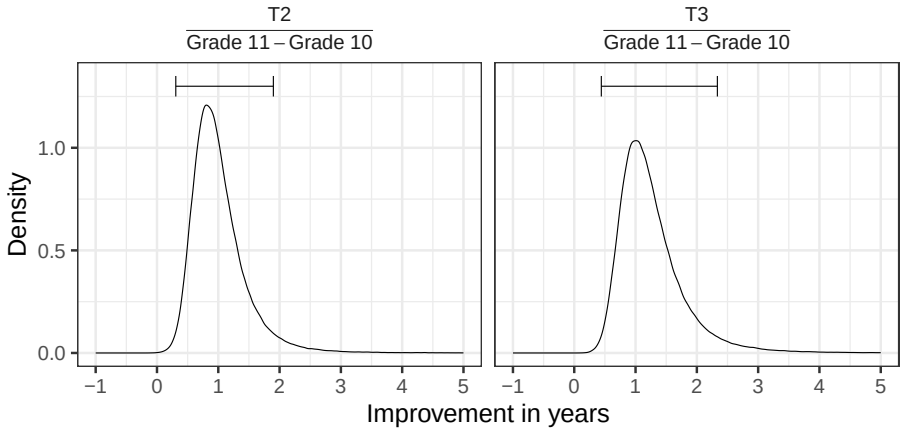

Figure 7: Improvement in the experimental study relative to the improvement between grade 11 and grade 10. The left panel shows the posterior distribution for the improvement between measurement 1 and measurement 2 divided by the improvement from grade 10 to 11 (95% HPD [0.31, 1.90]). The right panel shows the improvement from measurement 1 to measurement 3, standardized in the same manner (95% HPD [0.44, 2.33]). The horizontal error bars above the densities represent 95% HPD intervals.

how probable it is that 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 (Hojat & Xu, 2004; Fan, 2001).

Comparing the distribution of task-effect 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. Another thing these tests 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 (McShane, Gal, Gelman, Robert, & Tackett, 2019; Benjamin et al., 2018).

We chose 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. In a frequentist analysis, it is also possible to do this so some extent. A distribution of effect sizes can be obtained by using the point estimates

Limitations

It is important to stress that if improvement exceeds between-task effect the results cannot be viewed as a causal relation. This is no different from "correlation is not causation"; since no randomized experiment with a control group takes place the results are correlational and must be interpreted as such. To assert a causal relation a control group is essential.

A key assumption 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 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 in practice by accounting for the hierarchical structure (e.g., see 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).

Don: ik weet niet zeker of dit argument helemaal klopt.

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 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 inference done in the paper are based solely on the posterior distribution, the influence of the prior distribution is likely negligible.

Don: We kunnen vertellen dat frequentisten dit ook kunnen doen met een puntschatting van variantie. Het is dan alleen niet duidelijk is hoe de standaardfout van die puntschatting gebruikt kan worden om de onzekerheid van de schatting te laten voortvloeien in de gesimuleerde effect groottes (en het negeren van de onzekerheid zal leiden tot een te nauwe verdeling van effect groottes en als gevolg te zekere uitspraken gegeven de data).

Recommendations

Prior information can enrich the statistical analyses and provide more insight into the data. Here, we outline several 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 from an experimental study is that the data are comparable. Whether the data are comparable hinges on the validity of the measurement instruments. It is

not so much important that the instruments measure what they are supposed to measure (the traditional definition of validity), but rather they should measure the same construct. If the baseline study measured different constructs than the experimental study, for instance, because they used different measurement instruments, then a comparison is unintelligible and thus meaningless. Thus, we recommend to use 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, researches 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.

The use of our method is subject to large scale assessments publishing their results. It is key that those studies either disclose the raw data, or that they publish the marginal posterior distributions of the parameters. If the results of large scale assessments are not available as benchmark, then it is inoperable to use them to inform the analysis of experimental studies.

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.

Don: Hier moeten we iets meer zeggen over de mogelijkheden die dit introduceert.

References

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., ... Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6–10.

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.

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

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*(1).

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.

Graham, S. E., & Harris, K. R. (2014). Conducting high quality writing intervention research: Twelve recommendations.

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.

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

Lakens, D. (2017). Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, *8*(4), 355–362.

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

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

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*, 406–419.

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (Eds.). (2018). Editorial: Bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review*, *25*, 1–4.

Zwarts, M., Rijlaarsdam, G., Janssens, F., Wolfhagen, I., Veldhuijzen, N., & Wesdorp, H. (1990). Balans van het taalonderwijs aan het einde van de basisschool. *Uitkomsten van de eerste taalpeiling einde basisonderwijs*.

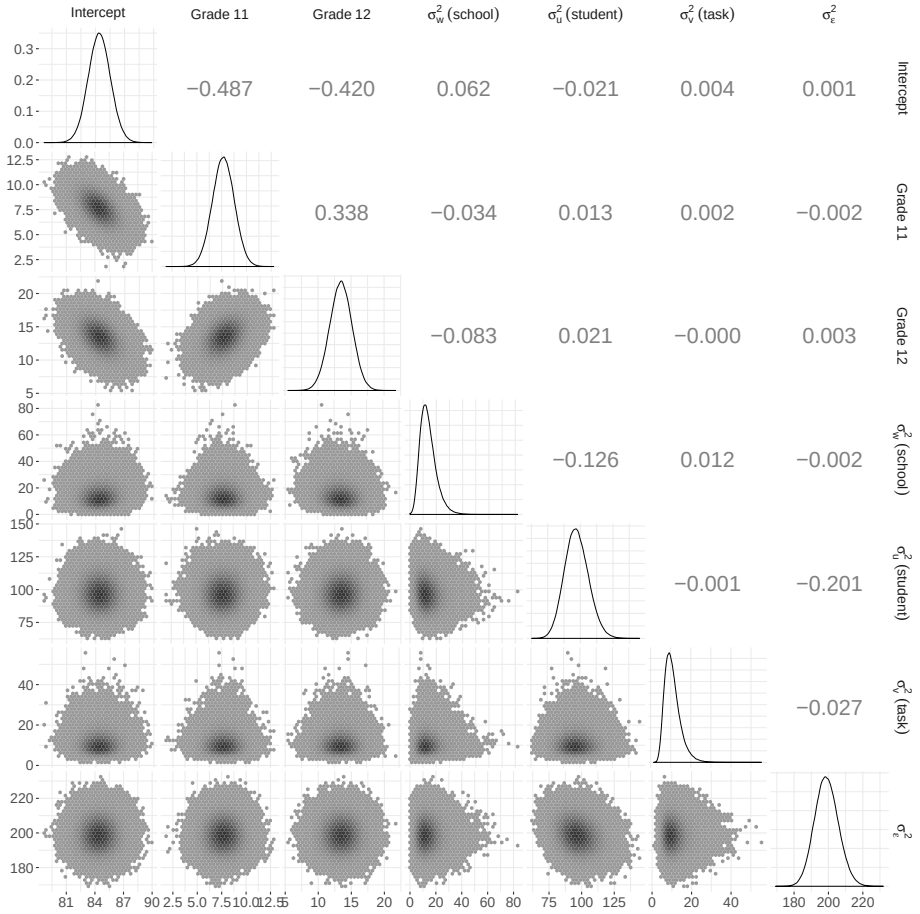

Figure 8: Visual summary of the posterior distributions for the group level effects of the baseline data set. The strips above and right of the figures indicate the parameters compared. Figures on the diagonal show marginal density estimates. Figures below the diagonal show bivariate hexagonal histograms. The numbers above the diagonal indicate the Pearson correlation between the samples of the parameters.

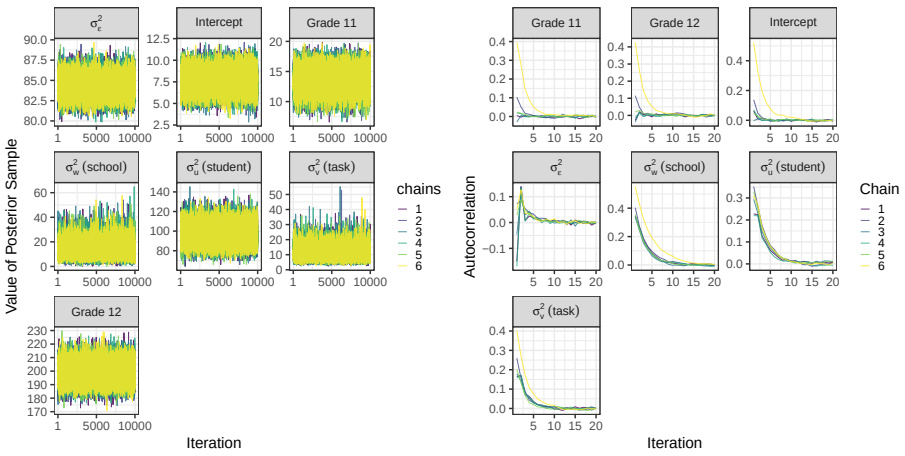

Figure 9: Convergence diagnostics for the analysis of the baseline data set. Left: trace plots of the first 10,000 posterior samples after warmup. The different chains appear indistinguishable, which indicates they converged. Right: Autocorrelation of the chains. The 0^{th} lag was omitted (as this is 1 by definition). The autocorrelation drops to 0 after about 5 iterations.

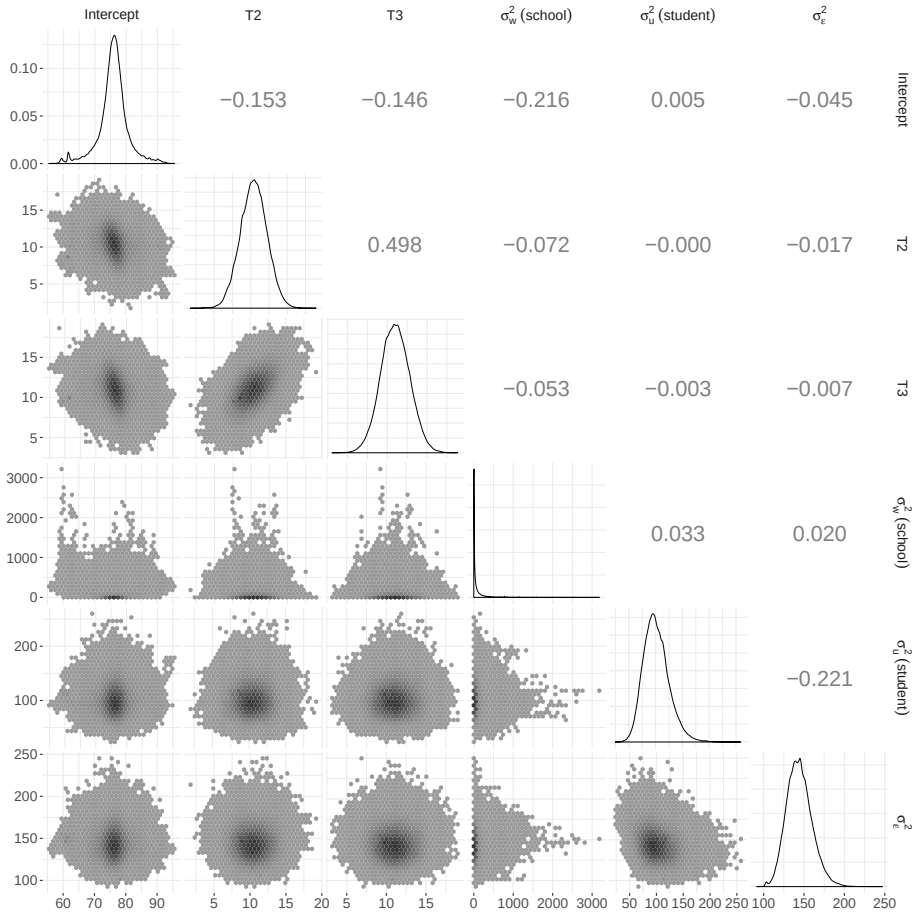

Figure 10: Visual summary of the posterior distributions for the group level effects of the experimental data set. The strips above and right of the figures indicate the parameters compared. Figures on the diagonal show marginal density estimates. Figures below the diagonal show bivariate hexagonal histograms. The numbers above the diagonal indicate the Pearson correlation between the samples of the parameters.

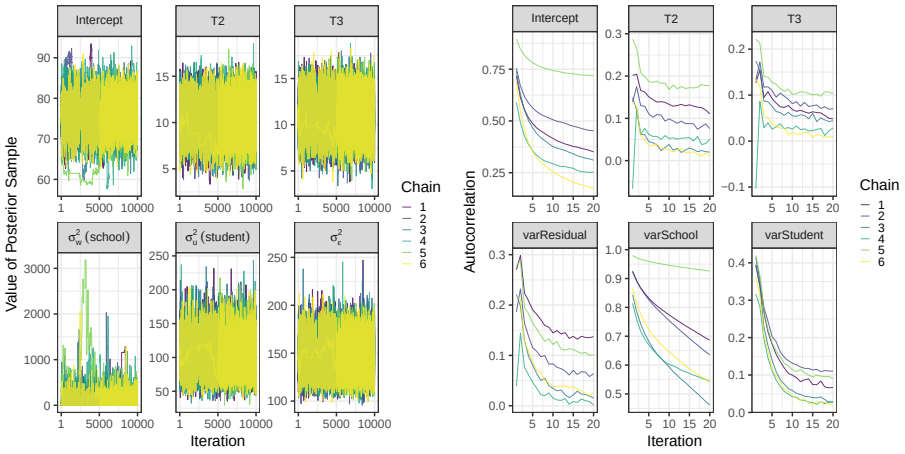

Figure 11: Convergence diagnostics for the analysis of the experimental data set. Left: trace plots of the first 10,000 posterior samples after warmup. The different chains appear indistinguishable, which indicates they converged. Right: Autocorrelation of the chains. The 0^{th} lag was omitted (as this is 1 by definition). The autocorrelation drops to 0 after about 5 iterations.