Informed Priors 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 . We adopt a Bayesian paradigm as this enables us to incorporate both the uncertainty in the estimates of a national assessment of the a Bayesian methodology for incorporating prior information from a

*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.

In many countries achievements of students is 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 in order 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.

A common denominator in national assessments is for all subject areas measurements are based on an analysis of that subject area. Therefore, students have to 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 (Xx). 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 prestest and one text as posttest (xx). Based on these texts we are proned to draw conclusions on changes in the writing skills of students. Hence, generalization over writing tasks does not appear to be a crucial issue in experimental writing studies as it is assessments. Although it is well documented that differences between different types of writing assignments can be large (e.g. ...), and we hardly can make inferences based on only one writing assignment (xx).

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 kind of control condition, or standard level of achievement for experimental studies. Second, results prior studies form the basis for new studies and research hypotheses. However, 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. May be the power could be increased if we would enrich our analyses with prior results. 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 manner 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 analyses are not equipped for that and a general approach seems far from obvious.

In this respect a Bayesian analyses of prior results and constructing approximate priors from the data might be preferable; Bayesian statistics offer a more rigorous and consistent approach to incorporate prior knowledge into statistical analyses. In Bayesian inference, prior knowledge is represented by probability distributions, which fully contains all uncertainty about the quantities of interest. Usually, prior distributions are chosen to be uninformative, i.e., such that they do not influence the results. However, in educational research there is an abundance of data and prior distributions are an ideal medium to propagate prior results into new analyses.

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 analysing the data with a multilevel model. We show how to approximate the posterior distributions with parametric distributions such that these can be used as prior distributions in future analyses. Next, we demonstrate the influence of these priors by analysing a follow up data set using both the newly obtained priors and uninformative priors.

Don: Some sentence about the diverence in results. The paper is concluded with a discussion on ... (todo).

Baseline Data Set: Lift...

Bayesian Inference

This section aims to give a brief introduction to Bayesian inference. 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).

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, imagine the intercept of

a regression model, β_0, is 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 to a posterior is given by Bayes theorem:

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

Here, the prior distribution of the parameters $\boldsymbol{\beta}$ 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 probably 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.

Bayesian Hypothesis Testing

Don: We moeten hier iets over zeggen, maar hoeveel?

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 obtain 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 mean of

the posterior distributions, we simulate many observations from the posterior distribution and use the sample mean of these observations to approximate the posterior mean. 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.

Extracting Priors from Posteriors

When updating the prior beliefs to posterior beliefs all observations are typically used at once. However, it is also possible to update the prior beliefs for a single observation and then use the resulting posterior distribution as a prior distribution for the next observations. That yields another posterior, which can then be used as a prior distribution for the third observation. A key aspect of Bayesian inference and philosophy is that the two approaches, updating all observations at once and updating one observation at a time, result in the same posterior distribution. This property shows that prior information can influence the Bayesian inference through the prior. In line with this is the statement of Lindley (1972, p. 2) "Today's posterior is tomorrow's prior".

National assessments that monitor student performance regularly provide insight in how students vary across assignments, classes, schools, and even years. However, knowledge about this variation is often not used in the statistical analyses by experimental studies. We aim to incorporate the information of the analysis of a baseline data set by describing the posterior distribution of a baseline data set

Exact Priors

Given a the data from a baseline study, X and the data from an experimental study, the posterior distributions may be written as:

$$p(\boldsymbol{\beta} \mid \mathcal{D}_X) = \frac{p(\mathcal{D}_X \mid \boldsymbol{\beta})\, p(\beta)}{p(\mathcal{D}_X)}$$
$$p(\boldsymbol{\beta} \mid \mathcal{D}_Y) = \frac{p(\mathcal{D}_Y \mid \boldsymbol{\beta})\, p(\beta)}{p(\mathcal{D}_Y)}.$$

If we use the posterior distribution of a baseline data set $p(\boldsymbol{\beta} \mid \mathcal{D}_X)$ as a prior for the experimental study we obtain:

$$p(\boldsymbol{\beta} \mid \mathcal{D}_Y) = \frac{p(\mathcal{D}_Y \mid \boldsymbol{\beta})}{p(\mathcal{D}_Y)} \frac{p(\mathcal{D}_X \mid \boldsymbol{\beta})\, p(\beta)}{p(\mathcal{D}_X)}.$$

The normalizing constant $p(\mathcal{D}_Y)$ is that which makes the density integrate to 1. Thus it equals

$$p(\mathcal{D}_Y) = \int_{\boldsymbol{\beta}} p(\mathcal{D}_Y \mid \boldsymbol{\beta}) \frac{p(\mathcal{D}_Y \mid \boldsymbol{\beta})\, p(\beta)}{p(\mathcal{D}_X)},$$

Here, $p(\mathcal{D}_X)$ is constant with respect to β, so we can write:

$$p(\mathcal{D}_Y) = \frac{1}{p(\mathcal{D}_X)} \int_{\boldsymbol{\beta}} p(\mathcal{D}_Y \mid \boldsymbol{\beta})\, p(\mathcal{D}_Y \mid \boldsymbol{\beta})\, p(\beta)\,.$$

Plugging this into the original expression for the posterior, $p(\mathcal{D}_X)$ cancels and the expression becomes:

$$p(\boldsymbol{\beta} \mid \mathcal{D}_Y) = \frac{p(\mathcal{D}_Y \mid \boldsymbol{\beta})\, p(\mathcal{D}_X \mid \boldsymbol{\beta})\, p(\beta)}{p(\mathcal{D}_Y)}.$$

Approximate Priors

In practice, the data from a baseline study may not be available (e.g., because of privacy concerns). Instead

Statistical Software

All analyses were done in R (R Core Team, 2019). The R package `brms` was used to generate the Stan code for the multilevel analyses (Bürkner, 2017). Stan is a probabilistic programming language for general purpose Bayesian inference that was used to obtain MCMC samples from posterior distributions (Carpenter et al., 2017). The R package `rstan` was used to interface between R and Stan (Stan Development Team, 2018). The R package `fitdistrplus` was used to approximate posterior distributions with parametric ones (Delignette-Muller & Dutang, 2015).

Baseline Analysis

Don: Wat willen we hier laten zien?

Application

(product) Data was collected from 276 students of two high-schools in the Netherlands. The students made three writing tasks in one week; one on a Monday, one on a Wednesday, and one on Friday. After the first task, the students received feedback on the quality of their written texts through ... .

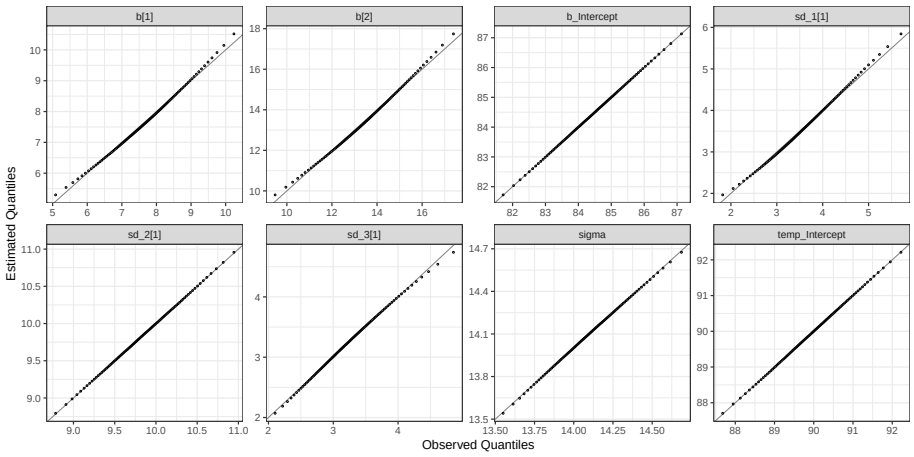

Figure 1: Quantiles of the MCMC samples (x-axes) against quantiles of the fitted gamma distributions (y-axis).

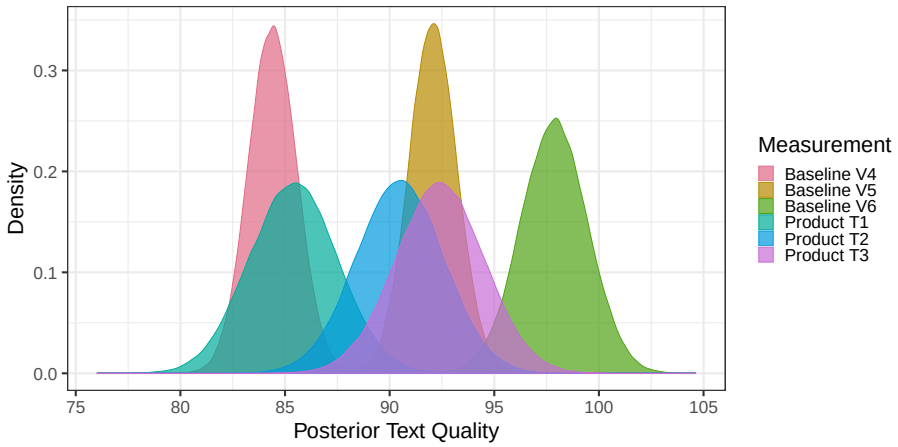

Figure 2: Posterior distribution of tekst quality in V4, V5, V6, and the three measurement occasions in ....

Discussion

Here, we introduced a procedure for incorporating prior information from large scale assessments into experimental studies. This

Recommendations

Prior information about a subject can enrich the statistical analyses and provide more insight into the data. However, we emphasize that a key requirement

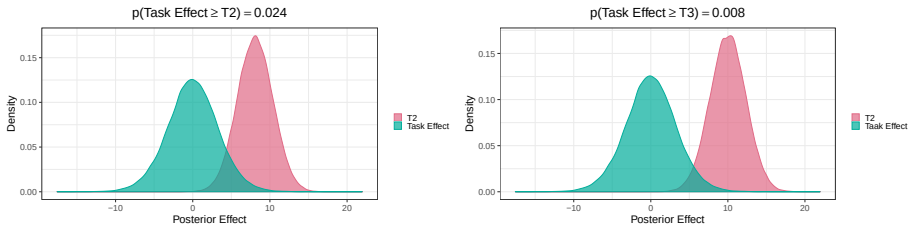

Figure 3: Posterior distribution of improvement versus

for comparison with a large scale assessment is that the data obtained in an experimental study are comparable with the data from a large scale assessment. Such a comparison hinges on the validity of the measurement instruments. If the constructs measured at baseline differ from those measured by a study then a comparison between them cannot be interpreted and is meaningless. Our recommendation is to use the same measurement instruments as those used in the baseline assessment.

Limitations

When we used parametric approximations to the posterior distribution of the baseline data set, we essentially used a three-step procedure. First, fit the baseline data set, second, second, fit parametric approximations to the MCMC samples, and third, use these parametric approximations in a subsequent analysis. However, the first two steps could be carried out simultaneously, i.e., the posterior distribution could be approximated directly by a parametric distribution. A benefit of this is that it likely reduces the approximation error because the MCMC samples are already a

in our procedure the numerical error of the

will propagates to the estimates. However, to assess the quality of the parametric approximation an unbiased estimate of the posterior is needed, which reintroduces the need for MCMC.

A second limitation of parametric approximations is that it can be complicated to model multivariate distributions. This is particularly important when the parameters are correlated. However, the parameters of interest in multilevel models, group variances, tend to be uncorrelated. This was also confirmed by the

In the example shown, this seemed unnecessary as the correlations among the posterior samples was

Altogether,

References

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).

Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, *64*(4), 1–34. Retrieved from http://www.jstatsoft.org/v64/i04/

Lindley, D. V. (1972). *Bayesian statistics, a review* (Vol. 2). SIAM.

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

Stan Development Team. (2018). *RStan: the R interface to Stan.* Retrieved from http://mc-stan.org/ (R package version 2.18.2)

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

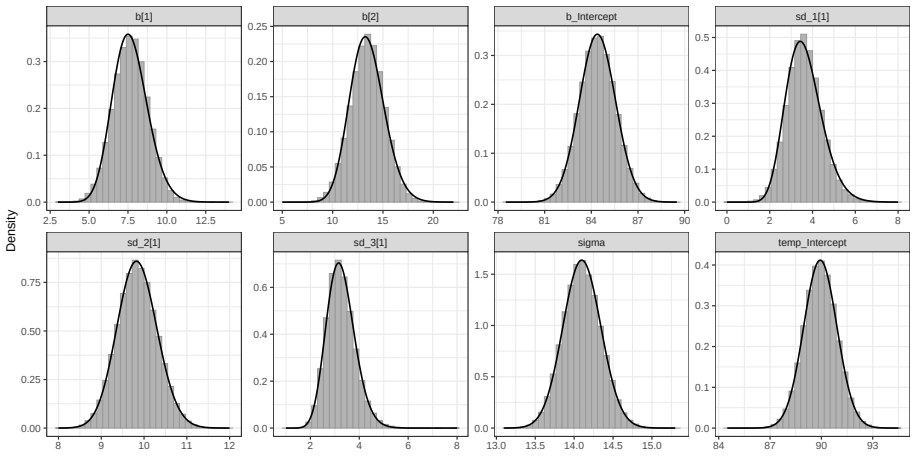

Figure 4: Histogram of posterior samples for the baseline data set. The black line overlayed represents a parametric approximation to the posterior distribution.

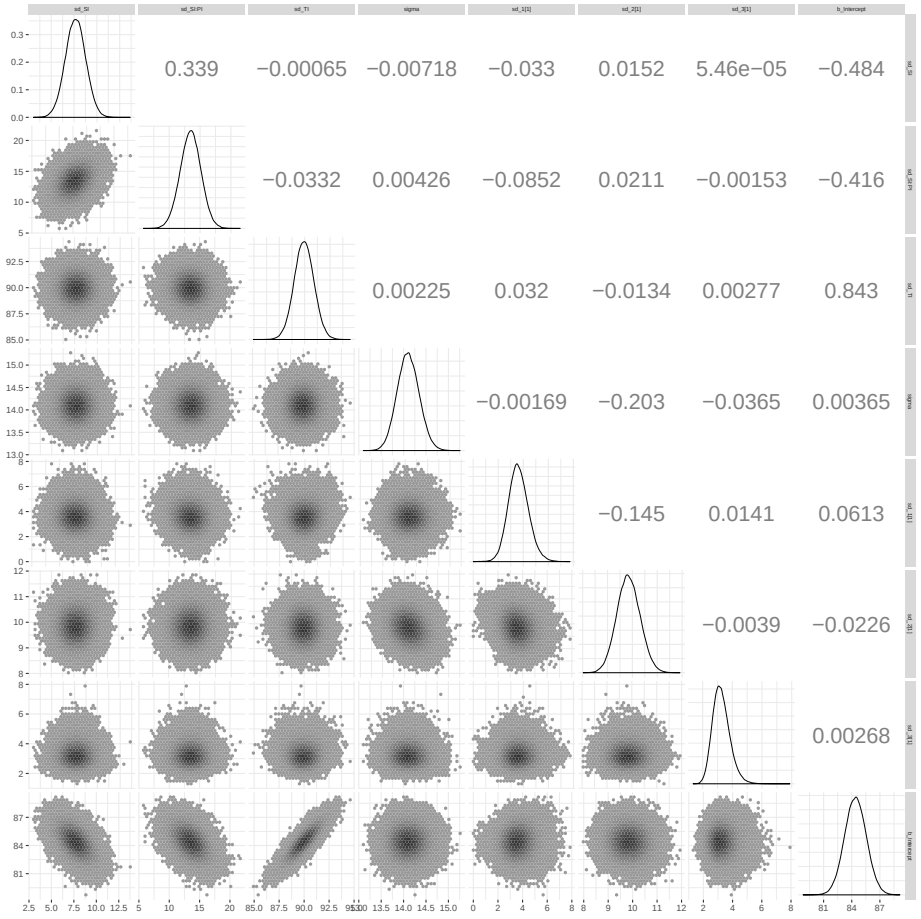

Figure 5: Visual summary of the posterior distributions for the group level effects of the baseline data set. The strips above and right of the figures indicate the parameters compared. Figures on the diagonal show marginal density estimates. Figures below the diagonal show bivariate hexagonal histograms. The numbers above the diagonal indicate the Pearson correlation between the samples of the parameters.