Overview of ‘Imagining the Sun’ analysis

The questionnaires

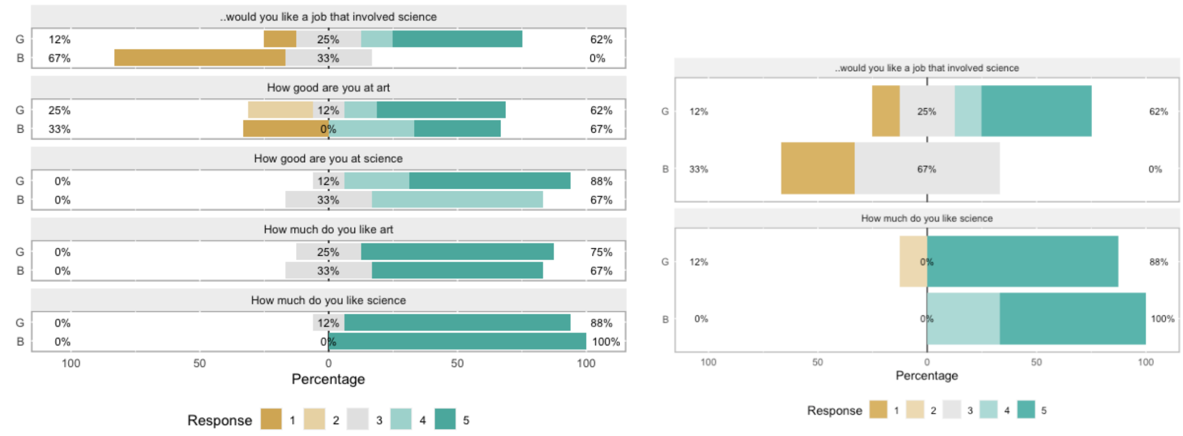

Figure 1 Example distributions of responses. *The figure shows the distribution of student response to a number of questions. The pre-activity responses are shown on the left and the post-activity responses are shown on the right. This is from a Primary School. Results are split into girls (G) and boys (B).*

The school children involved in the outreach project were given questionnaires which asked them about their opinions on several statements. They were able to rate their feelings on a scale that went from of ‘strongly disagree’ to ‘strongly agree’ with 5 intervals.

An example of summary data from the survey responses from a primary school is shown in Figure 1, where we have relabelled the categories from 1 (‘strongly disagree’) to 5 (‘strongly agree’). The original scale presented to the children was a series of faces with expressions from sad to happy with textual indicators to the sentiment, i.e. ‘*not very good*’ or ‘*not at all’* for the lowest rank category, and ‘*very good’* or ‘*lots*’ for the highest rank category. This representation can be classed as a Likert-type Scale. We note that the numerical labels assigned by us in the summary plot infer no information about the nature of the intervals between the categories/levels, i.e., there is no guarantee that interval between the categories is equally spaced.

Working with Ordinal data

Likert-type scale data are ordinal, meaning there are a discrete number of categories/levels which have a clear ordering, but the distance between categories/levels is not known. The category widths may also not represent equal increments.

A typical approach to analysing Likert-type data of this form is to assume that the data is metric, and apply metric methodology, e.g., taking the mean of the scores. However, it has been suggested that applying metric models to ordinal data can lead to misinterpretation (Liddell & Kruschke, 2018). As such, we follow the suggestion of Liddell & Kruschke, and analyse the data with an ordinal model, in this case an ordered-probit model.

What is the difference between the metric and ordinal models? Here, the categories 1 to 5 represent responses to a question on a Likert-type scale. A metric model assumes that the probability of a response is the normal probability density at that response value. This approach inherently assumes the distance between responses is equal. The ordered-probit model, however, assumes that the probability of a response is the cumulative probability between two thresholds on an underlying latent continuum. For the ordered-probit model, the latent continuum is considered to be a Normal distribution with some mean and standard deviation. Determining the latent parameters is referred to as ordinal regression.

Analysis of data

In the following we focus on two of the questions that we believe can inform us about the career aspirations of the students, namely i) *Would you like a job that involved science*?; ii) *How much do you like science?.* We are interested in how the response to these questions changes after the students have had an interaction with a researcher.

The model

To estimate the change in response, we model the pre and post survey answers to each question with an ordered-probit model. The model assumes that the latent distribution is a normal distribution, $\mathcal{N}(\mu, \sigma)$, and requires estimating the mean, μ, the standard deviation, σ, along with the threshold values between the intervals, θ_K (with K thresholds). Due to the nature of the problem, if μ, σ, and θ_K are be modelled jointly, there is then non-identifiability in the observational model (i.e., there is no unique solution). One way to navigate this problem is that two of the parameters must be fixed to remove the degeneracy. Two options are to: *i)* fix the mean and the standard deviation of the latent normal; or *ii)* fix the outer thresholds and estimate the interior thresholds (of which there are $K-2$).

Here, we choose to fix the outer thresholds. The number of thresholds used varies for each question and depends upon the response categories/levels containing data. For example, there are responses at each of the 5 levels for the data shown in Figure 3, hence there are four thresholds used ($K=4$), with two fixed at the $\theta_1 = 1.5$ and $\theta_4 = 4.5$. For the data shown in Figure 6, the post-workshop survey has four levels with data, hence we use three thresholds and $\theta_1 = 2.5$ and $\theta_3 = 4.5$. An alternative approach is to come up with a principled prior model that can regularise the cut-points, which avoids fixing model parameters (see Betancourt 2019).

To estimate the model parameters, we follow a Bayesian approach as outlined in Kruschke (2015). We are interested in calculating the posterior distribution, $p(\vartheta|y)$, which describes probability of the parameters, ϑ, (i.e., the mean, standard deviation, and interior thresholds) given the data, y. To obtain the posterior, we require a likelihood function and prior distributions for the parameters.

The likelihood function gives the probability of outcome k given fixed model parameters, ϑ:

$$p(y=k|\vartheta) = \Phi\left(\frac{\theta_k - \mu}{\sigma}\right) - \Phi\left(\frac{\theta_{k-1} - \mu}{\sigma}\right),$$

where Φ is the cumulative distribution function for the normal distribution. We supplement the likelihood with the prior distributions for the mean, standard deviation, and threshold values. We again follow Liddell & Kruschke, using weakly informative priors. For the latent mean parameter, we specify a normal prior density function:

$$\mu \sim \mathcal{N}((K+1)/2, K),$$

with the prior mean being the midpoint of the ordinal scale and standard deviation equal to the number of ordinal categories.

For the latent standard deviation, we set a uniform prior density,

$$\sigma \sim \mathcal{U}(0,10K).$$

And for the threshold values, we set a normal prior density,

$$\vartheta_k \sim \mathcal{N}(k+0.5, 2) \text{ for } k = 2, K-2.$$

For each question, we construct the model such that the interior thresholds are the same for both pre and post responses but let the latent mean and standard deviations differ. This assumes that the respondents did not change their perception on the intervals between levels in the time between the pre and post questionnaires. Due to small samples sizes (and only 2 groups per education level) we group together responses from the different schools. An alternative option would be to undertake a hierarchical approach, modelling the parameters for each school as being distinct but drawn from common population distributions (see, e.g., Gelman et al. 2014).

Estimating the posterior distribution

To obtain the joint and marginal posterior distributions for the parameters, we employ Markov-Chain Monte Carlo (MCMC) sampling, using the MCMC sampler JAGS (Plummer 2003). The quality of the samples drawn are analysed with the standard methods, namely through calculation of the auto-correlation coefficients of the sample trace and the Gelman-Rubin $\hat{r}$ statistic (see, e.g., Kruschke 2015; Gelman et al. 2014). There is no indication of convergence issues for the sampling.

Results

Before examining the model results, we should examine whether the model is able to accurately describe the data. One such method of doing this is to compare the posterior predictive distribution to the data, i.e., a posterior predictive check (e.g., Gelman et al. 2014). The posterior predictive distribution is the distribution of unseen values, conditioned on the values we have already seen. Any obvious discrepancies between the posterior predictive distribution and the distribution of our data would call into question our modelling choices. Figure 2 shows an example of the data for a question and the posterior predictive distribution for each response, displaying the posterior mean value of each response category and the 95% Highest Density Interval (HDI). As can be seen, there is good agreement between the observed and predicted distributions.

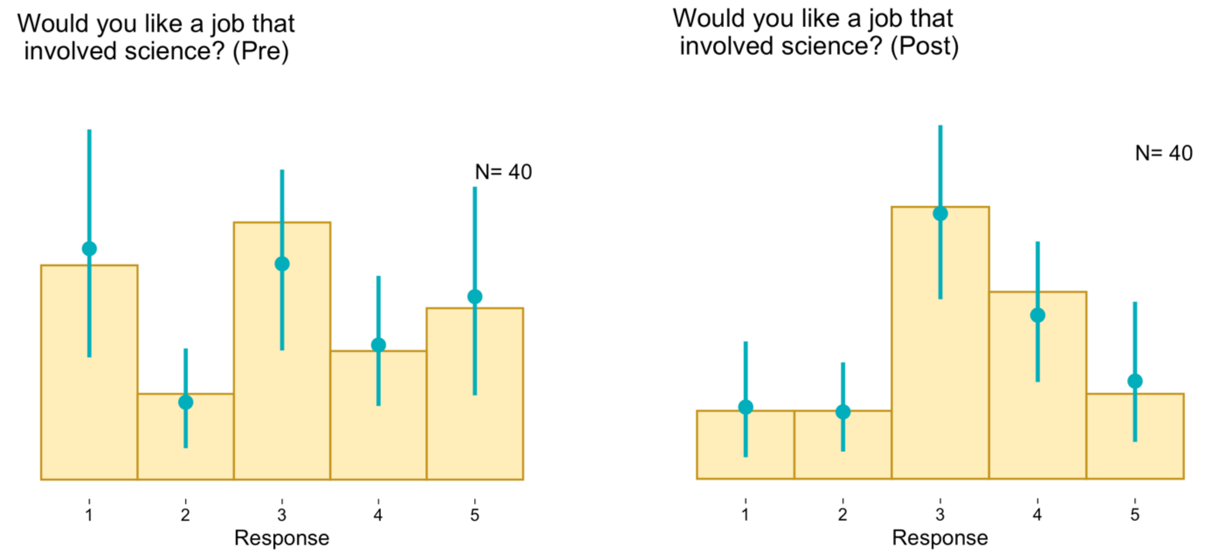

Figure 2 Data distribution and posterior predictive distributions*. Primary data (Would you like a job the involved science?). Left is the distribution of the responses for the pre survey (yellow) and posterior predictive means (blue dots) and 95% HDI (blue lines). Right is for the post survey.*

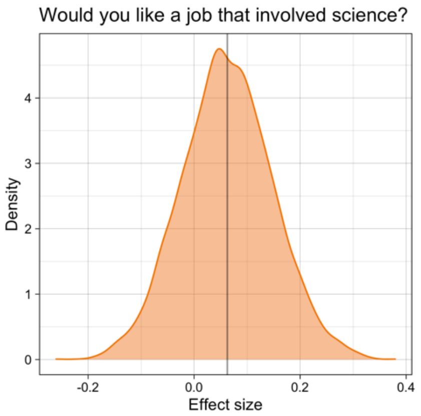

Figure 3 Posterior distribution for effect size. *The orange curve and shaded region represents the posterior density and the black vertical line is the mean effect size.*

In terms of the examining the impact made on the children's attitudes, we are not so much interested in what the posterior distributions of parameter values (e.g., the latent mean, etc.) are for each set of observations. Of real interest is the difference between opinions after the intervention. One way that this can be assessed is by looking at the effect size:

$$effect\ size = \frac{\mu_{post} - \mu_{pre}}{0.5\sqrt{\sigma_{pre}^2 + \sigma_{post}^2}},$$

which is the difference between means scaled by the pooled standard deviations (this is known as Cohen's d). In this calculation we use the samples from the joint posterior distributions of latent mean and standard deviation to calculate the posterior distribution for the effect size, which is shown in Figure 3. The mean of the effect size is larger than zero, which would suggest that there has been a positive shift in the children's attitudes towards whether they would like a job that involved science. However, the mean value is <0.1, which can be classed as a small effect. Furthermore, it can be seen there is a significant probability associated with a null/negative result. This means that we cannot be certain about the effect size and that there has been a positive change in attitudes.

In Figure 4 we show similar results for the question *'How much do you like science?'*. The posterior predictive checks again demonstrate that the ordered probit model provides a reasonable description of the observed data. The right-hand panel displays the posterior distribution for the effect size, and the mean of the effect size is positive but very small - close

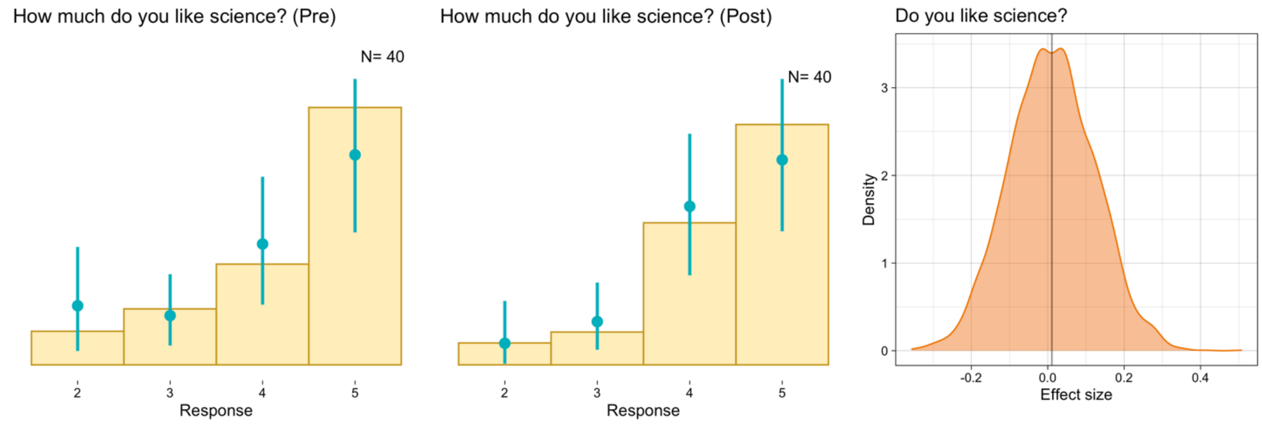

Figure 4. Results for '*How much do you like science?*' (Primary). The left and middle panel show the distribution of responses (yellow) and the posterior predictive distributions for the ordered-probit model (blue points and lines). The right-hand panel shows the posterior distribution of the effect size.

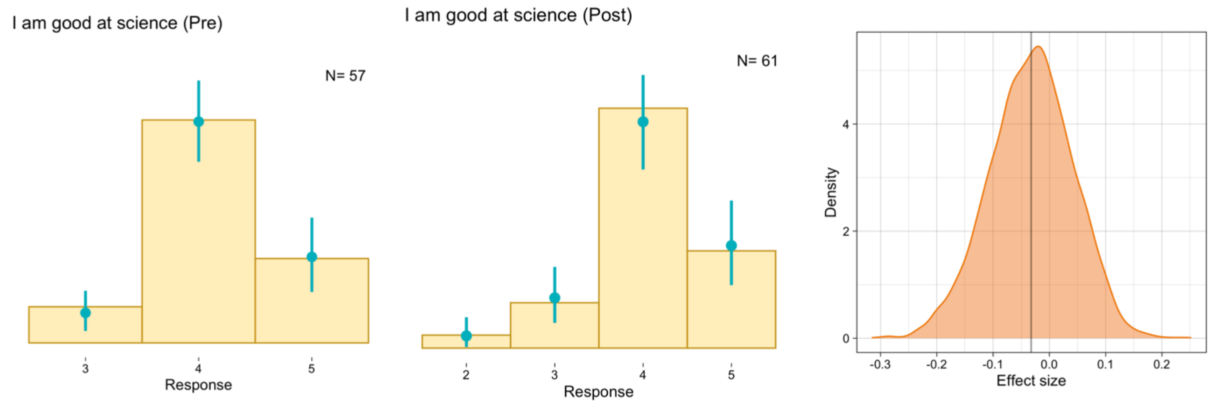

Figure 5. Results for '*I am good at science?*' (Secondary). The left and middle panel show the distribution of responses (yellow) and the posterior predictive distributions for the ordered-probit model (blue points and lines). The right-hand panel shows the posterior distribution of the effect size.

to zero. As with the previous question, this suggests the effect size is small but the uncertainty in the value is too large to make meaningful statements about its actual value.

In Figure 5 we show similar results for the question *'I am good at science?' (Secondary).* The posterior is similar to our previous examples, although in this case the mean of the effect size is negative but small.

No More Marking

In addition to the questionnaire, the students were also asked to 'W*rite or draw what the Sun makes you think about'*. We wanted to measure whether the students would include a greater amount of scientific content in their answers after the intervention. In order to assess the results, No More Marking was used. The tool enables comparative judgement of pieces of work. The data returned from the tool are `quality` scores, based on a Rasch logistic model (Bradley-Terry-Luce model, see. Wheadon et al., 2020). Both pre and post responses were assessed together, meaning the scores provided the relative `quality` of all responses (as opposed to relative quality of either pre or post responses alone). The Bradley-Terry-Luce model assigns scores around a zero mean, which were rescaled between 0 (lowest content) and 100 (highest content).

Using the No More Marking results, we take the scaled scores and calculate the difference between the pre and post scaled scores for each student. We are interested in whether there is positive change is scores for the students. One way to assess this is to frame the differences between scores as a binomial problem. An increase in a quality score for an individual student can be classed as a success, while a decrease in score is classed as a failure.

One could argue that, should there be no impact from the intervention, then the number of successes and failures should be binomially distributed with the relative frequency of success being 50%, i.e., we should expect roughly equal numbers of positive and negative quality score differences. If there is a positive impact, then the relative frequency for success should be greater than 50%. A null hypothesis for such a test could be $H_0: \theta = 0.5$ and our alternative hypothesis is $H_1: \theta \neq 0.5$.

To examine the relative frequency of success, we undertake a Bayesian analysis of the differences between post and pre quality scores.

For the likelihood we use is a Bernoulli distribution, with a probability mass function:

$$p(k) \sim \theta^k (1-\theta)^{1-k},$$

Where θ is the relative frequency of a success ($k = 1$) and $(1-\theta)$ is the relative frequency of a failure ($k = 0$). We put a beta prior on the probability of success, with the shape hyperparameters set to $\alpha = 2, \beta = 2$, hence this is a weak prior centred on $\theta = 0.5$.

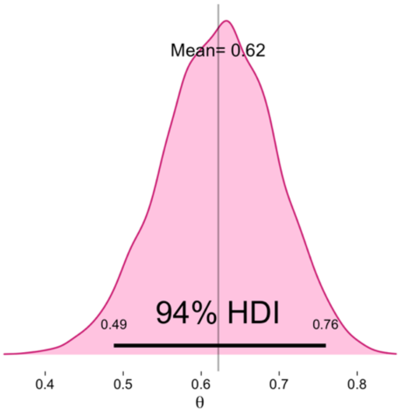

Figure 6. *Posterior distribution for relative frequency of success, ϑ. The vertical line shows the mean value, while the horizontal line indicates the Highest Density Interval (HDI), i.e. the region that contains 94% of the probability mass.*

The posterior distribution for θ is shown in Figure 6. The mean value is 0.62, which suggests that there has been an increase in the scientific content in the pictures/words used by the students. However, the posterior distribution for θ is broad and covers $\theta = 0.5$, with ~5% of the probability mass at or below $\theta = 0.5$. Hence, we are not able to confidently rule out the possibility that the intervention had no influence on the students.

A hypothesis test can be carried out to examine whether there is evidence against the null, using Bayes factors (namely the Savage Dickey ratio, see, e.g., Wagenmakers et al., 2010). Such a test confirms that, given the current results, there is not strong enough evidence to reject the null.

References

Betancourt, M. (2019) An Ordinal Regression case study https://betanalpha.github.io/assets/case_studies/ordinal_regression.html

Gelman, A. et al., (2014) Bayesian Data Analysis (3rd Edition) CRC Press, Chapman and Hall

Kruschke, J. K. (2015) Doing Bayesian Data analysis: a tutorial with R, JAGS, and Stan (2nd Edition), Academic Press, Elsevier

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology, 79,* 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

Plummer, M. (2003) JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling *Proceedings of the 3rd international workshop on distributed statistical computing*

Schäfer, T. and Schwarz, M. A. (2019) The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases, Front. Psychol., Volume 10.

Wheadon, C., Barmby, P., Christodoulou, D. & Henderson, B. (2020) A comparative judgement approach to the large-scale assessment of primary writing in England, Assessment in Education: Principles, Policy & Practice, 27:1, 46-64, DOI: 10.1080/0969594X.2019.1700212

Wagenmakers, E. J. et al., Cognitive Psychology 60 (2010) 158–189