Bayesian item-response theory models for educational assessment data

Robert L Grant

BayesCamp Ltd
16 City Business Centre, Hyde Street,
Winchester, SO23 7TA, United Kingdom

February 15, 2022

Abstract

Bayesian statistical models provide a flexible way of analysing data on educational assessments. This paper introduces the concepts, using a probabilistic programming language to set out the models in a human-readable way, and building complexity one step at a time. They allow us to account for student background differences, progress over time, resits, varying course difficulty, missing data and results that are only recorded as pass/fail or grades. Three case studies, originally analysed by the author, serve to illustrate additional twists or practical considerations. Options for communicating findings are discussed, and reflections on the practicalities for researchers and managers close the paper. This method is preferable in several ways to a traditional approach using classical regression and hypothesis testing.

Keywords: tertiary education, assessment, attainment, student participation, bias, Bayesian statistics.

Contents

1 Introduction

The purpose of this paper is twofold: to promote flexible Bayesian item-response theory (IRT) models for educational assessment data, and to record in the public domain the precise methods used in three previously published analyses.

The methods set out here apply to research that uses the results of assessments in an education setting. I will refer to these generally as "marks". These might provide percentage marks, grades, or binary pass/fail data, or a mixture of these. The objective might be to compare marks between discrete times, or between classes, or between interventions. It might also take a regression approach, evaluating relationships between independent variables such as the amount of instructor contact time and marks (the dependent variable).

Any analysis must be crafted to suit the input (data and research question) and the output (the audience, their statistical literacy and their intended decision-making). We begin by defining the input, then the models, then the output. Then, we examine three case studies in detail. Code is given based on the Bayesian software BUGS[1] as it is easily read by humans, but these three case study projects were originally analysed using Stan,[2] and links to the original Stan code are given.

This paper assumes an understanding of the very basics of Bayesian modelling*, and of statistical analysis in general, up to the level of a classic Snedecor-inspired introductory course, which is to say the level that most education researchers encounter, in my experience. This includes linear regression and hypothesis testing.

2 Input: the data

A relevant dataset will contain variables such as these:

- marks
- the student ID relevant to each mark
- whether the mark is a percentage or grade or pass/fail
- the course or assessment ID relevant to each mark

*For a primer, I recommend the YouTube videos of Rasmus Bååth; those who want to learn more can follow up with those of Richard McElreath, and either *The BUGS Book*[1] or *Statistical Rethinking*.[3]

- optionally, other information about the delivery of the education, such as the instructor or year
- if you are evaluating an intervention, which "arm" of the study this mark belongs to
- other variables, if you are evaluating relationships with them, such as the student's previous qualifications or first language
- whether the assessment is a resit; perhaps the number of the resit

The mark obtained by a certain student, taking a certain course, is likely to be affected by factors of interest to educational researchers, such as the student's previous qualifications, the amount of contact time with tutors in the course, or the time of day of the assessment. These predictor variables may relate to the student, the course, or the specific assessment. When students who fail an assessment are able to re-take it, this adds another layer of detail.

These aspects may impact the analysis even if they are not the primary focus of the analysis. The most common form that this takes is called adjusting for confounding. Suppose that your institution brought in new technology for remote group tutorials, and you are tasked with evaluating it by comparing marks before and after its implementation. One might imagine that a simple t-test or Mann-Whitney would suffice*.

However, students are not randomly allocated to the technology; rather, it was implemented campus-wide. So, you must consider what else might have changed at the same time. Suppose the country plunged into economic recession at the same time. Now, two things have changed: the new technology has some effect, and also students from poorer backgrounds are less likely to apply, which means that the post-intervention student body has greater access to external support, private tutors, and so on. If you see an improvement, you must try to separate the economic advantage from the intervention's effect.

2.1 Simple linear regression

A very basic regression model might take this form:

$$\hat{y}_i = \beta_0 + \beta_1 z_i + \beta_2 x_i \tag{1}$$

*Readers from a commercial data science background may call this A/B testing.

where $\hat{y}_i$ is the predicted mark (percentage) obtained by the ith student. x_i is a binary variable (taking values 0 or 1) which is 1 for those assessments after the intervention. z_i is a measure of student i's socio-economic status. We are assuming for now that each student only appears once in the data.

The βs are regression coefficients that the computer must estimate for us. A one-point increase in socio-economic status leads to a β_1 increase in marks, on average. Then, the estimated effect of the intervention is β_2.

There is just one more aspect to add to this model: the observed marks, y_i, are not equal to the predicted values $\hat{y}_i$, but vary above and below it by chance[†]. We can model this with a probability density function. In this paper, we will restrict ourselves to the normal distribution, but I will comment on other distributions in the Discussion. We simply add a second line to our model:

$$\begin{aligned} \hat{y}_i &= \beta_0 + \beta_1 z_i + \beta_2 x_i \\ y_i &\sim \text{Normal}(\hat{y}_i, \sigma) \end{aligned} \tag{2}$$

The first line is *deterministic*: given x_i, z_i, and the β coefficients, the predicted mark $\hat{y}_i$ is perfectly determined. The second line is *probabilistic* or *stochastic*: it does not determine the exact value of y_i, but tells us how probable or improbable a value is, given $\hat{y}_i$ and the residual standard deviation σ, which describes the extent of scatter above and below the prediction.

So far, we have set out a simple linear regression model. Because one of the predictors is binary and the other continuous, you might recognise it as equivalent to ANCOVA, depending on the sort of statistical training you have had in the past. However, it pays to conceive of all these models in a regression framework, because then it is simple to extend them to your needs.

I have described this model as pertaining to a project that evaluates the impact of the intervention and must adjust for socio-economic status, but it would just as well apply to a project that evaluated the impact of recession, and had to adjust for this unfortunately coincident technology change. In that case, you would simply pay most attention to the other β coefficient.

[†]By "chance" or "random error", we mean stuff we are not interested in, and are willing to ignore in our model. The fact that student i had to move house mid-course and missed a couple of lectures might simply be regarded part of a large collection of exigencies, some increasing and some decreasing the observed marks. When they are added together, they may well create a normal distribution, thanks to a very useful statistical fact called the Central Limit Theorem.

2.2 Multilevel regression

In practice, we do not have a stream of unrelated marks; they belong to certain courses and certain assessments. Some assessments tend to produce higher or lower marks on average. What if, post-recession, most students played it safe by taking "easy" courses known to have given higher marks on average in the past? There would be a new source of confounding from the course.

We cannot adjust for the courses in the same way as we did the socio-economic status, because they have changed the probabilistic line in the model. Instead of each line in the data being scattered according to the same distribution, we must now think about the between-course scatter and the within-course scatter. (Again, you might be reminded here of ANOVA and its variants; you would be right, but again, we will use a regression framework because ANOVA only gets you this far and no further.)

Our model now has another line to accommodate a parameter v_j, which is sometimes called a "random effect". The subscript j indicates a course, so there is a v_j for each course. Also, the mark now becomes y_{ij}, for student i in course j. The intervention x_{ij} also has this, but the socio-economic status z_i applies to the student alone. Some are positive, increasing that course's marks over the overall average, and some are negative, decreasing them. We need to model how these v_js are scattered, so there is a probabilistic line for them:

$$\begin{aligned} \hat{y}_{ij} &= \beta_0 + \beta_1 z_i + \beta_2 x_{ij} + v_j \\ y_i &\sim \text{Normal}(\hat{y}_i, \sigma) \\ v_j &\sim \text{Normal}(0, \tau) \end{aligned} \tag{3}$$

The course random effects v_j are distributed according to another normal distribution, around zero this time, with standard deviation τ. This model's σ captures only the within-course standard deviation, and it is worth noting that it assumes that all courses have the same σ.

This is a multilevel model, and it requires a different algorithm in the software to estimate the βs. To help you recognise what is going on in your software output, these algorithms include names like restricted maximum likelihood (REML) or restricted iterative generalized least squares (RIGLS).

It may seem tempting, with data like these, to summarise each course's results into a mean (for example), and then analyse those means. However, in general, not every student takes every course, and so one would not quite be comparing

like with like. I have heard this called aggregation bias, or the aggregation fallacy, though not very often.

2.3 Item-response theory, or Rasch, or crossed random-effects models

Now, we must add one more complication: not only do marks belong to particular courses, but also to particular students. We need a student random effect alongside our course random effect:

$$\begin{aligned} \hat{y}_{ij} &= \beta_0 + \beta_1 z_i + \beta_2 x_{ij} + u_i - v_j \\ y_i &\sim \text{Normal}(\hat{y}_i, \sigma) \\ u_i &\sim \text{Normal}(0, \tau_u) \\ v_j &\sim \text{Normal}(0, \tau_v) \end{aligned} \tag{4}$$

This model deals with normally-distributed data. In psychometrics, the same structure with binary data is often called an item-response theory (IRT) model. We can imagine its use in the 1970s to develop occupational aptitude tests, if i identifies the respondent and j the question (item). It is traditional for the v_j term to be *subtracted*, as seen above. Of course, it makes no difference to the solution, as it merely serves to reverse the signs of the v_j values. We can describe u_i as student ability and v_j as course difficulty.

In biostatistics and social science, it is sometimes called crossed random effects* (we can picture each row in the dataset being linked to a student and to a course), and in psychology it is often called a Rasch model. Proponents of these approaches are characterised by infighting over their precise meanings, boundaries and ontologies; they may well object to being grouped together as one concept. However, the structure of the model as set out above is common, and this is what we will use.

These models can be estimated or "fitted to the data" using either frequentist or Bayesian methods. In the frequentist approach, the software starts from a guessed initial value for the β coefficients, and iteratively maximises their "fit" to the data, contingent on the values of σ, τ_u and τ_v. It does not have to estimate the actual values of all the u_is and v_js; in fact the philosophy of frequentism forbids that.

*Some statisticians, notably Andrew Gelman, don't use the term random effects, as the effects are not really random (though they are ignored in some software, in the same sense that we ignore random "noise", explained above). We might prefer "crossed group-level effects" models, but here I shall simply call it IRT.

3 The basic Bayesian models

Bayesian approaches, where probability distributions can be used to represent any source of uncertainty, allow for a different approach, directly estimating the u_is and v_js. This is helpful for checking the analysis, and allows outputs such as course league tables, incorporating uncertainty about ranks. Along with the βs, we begin with a prior distribution and update that using the data to arrive at a posterior distribution, whence we make all our estimates and inferences.

On the face of it, all we need to do is to add probabilistic lines to our model, representing the priors. We should note that these are marginal priors — there is one for each parameter — but in reality we update a single *joint* distribution over all the parameters, allowing for correlations between them. Marginal priors simply set those prior correlations to zero.

$$\begin{aligned} \beta_0, \beta_1, \beta_2 &\sim \text{Uniform}(-30, 30) \\ \sigma, \tau_u, \tau_v &\sim \text{Uniform}(0.01, 25) \\ \hat{y}_{ij} &= \beta_0 + \beta_1 z_i + \beta_2 x_{ij} + u_i - v_j \\ y_{ij} &\sim \text{Normal}(\hat{y}_{ij}, \sigma) \\ u_i, v_j &\sim \text{Normal}(0, \tau) \end{aligned} \tag{5}$$

There is one deterministic line, and four probabilistic ones. Of those, one refers to known data on the left hand side (y_{ij}), which specifies what we call the likelihood, the match between the model and the data, while the others refer to unknown parameters and specify the priors.

Bayesian analysis of these models is performed by repeated computer simulation. The most common method is called Markov chan Monte Carlo (MCMC). Initial values are set for the parameters, and the densities are calculated for each of the probabilistic lines. Multiplied together, these give the posterior density. The software "moves" to a nearby set of parameter values, and does the same, comparing the new posterior density with the old one. If the new one is better, it probably stays there and repeats the process; if not, it probably goes back and duplicates the previous values.

Repeating this many times gives us a *sample* from the joint posterior, perhaps containing a few thousand *draws* of parameter values. We can then estimate the parameter values by, for example, simply calculating the means. We can also assess the uncertainty around those estimates by, for example, finding the quantiles of the

posterior sample which give the central 95% of the probability — a set of credible intervals.

3.1 Probabilistic programming languages

In practice, we do not write out the mathematical formulas seen above* but instead use a similar computer language. These are collectively "probabilistic programming languages" (PPLs), although they are not used for general-purpose programming like C++. They just control one specific task, and are usually part of a bigger script written in R, Python, Stata, and the like. In this paper, I will present models in a pseudocode similar to BUGS[1] and JAGS.[5]

Our first Bayesian model in Equation 5 can be written in a PPL like this:

```
// priors:
beta_0 ~ uniform(-30, 30)
beta_1 ~ uniform(-30, 30)
beta_2 ~ uniform(-30, 30)
sigma ~ uniform(0.01, 25)
tau_u ~ uniform(0.01, 25)
tau_v ~ uniform(0.01, 25)
for(i in 1:n_students) {
   u[i] ~ normal(0, tau_u)
}
for(j in 1:n_courses) {
   v[j] ~ normal(0, tau_v)
}

// likelihood:
for(k in 1:n_marks) {
   yhat[k] = beta_0 + beta_1*socio_economic[k] +
             beta_2*intervention[k] +
             u[student_id[k]] -
             v[course_id[k]]
   y[k] ~ normal(yhat[k], sigma)
}
```

*If you like working with them, you may enjoy the MLwiN software from the University of Bristol,[4] which allows algebraic specification of multilevel models, including MCMC and RIGLS in its methods.

Mostly, this maps directly to what we have seen in algebra. There are two differences we must clarify before moving on. We use loops like `for(j in 1:n_courses) { ...}` to evaluate all the course effects from `1` to `n_courses`. We also have `n_marks` rows in our data, and each row has a `student_id` number, and likewise a `course_id` number. So, in the likelihood, for each row, we obtain the student and course numbers, and plug those into `u[ ]` and `v[ ]`.

In my pseudocode, `uniform()` distributions take the minimum and maximum values as the inputs or "arguments" inside the brackets. `normal()` distributions take the mean and standard deviation (some software might require the variance, which is the standard deviation squared, and others, including BUGS and JAGS, require the precision which is the reciprocal of the variance).

For the rest of this paper, I will specify models in this pseudocode, showing any lines that are changed in `teal coloured type`.

I will not talk further about the choice of priors here, except to note that they are examples of "weakly informative priors" (WIP): they guide the computer away from impossible values (there is just no way that any binary intervention to improve student performance is going to raise marks by more than 30% on average) and computationally troublesome values (sometimes, extremely small standard deviations can cause problems*).

3.2 Resits

When students resit assessments, there are multiple marks for the same combination of student and course. We can reasonably expect them to be independent of each other, given the student ability, the course difficulty and the order of attempts. We already have the first two in the previous model, and must add in the order of attempts. The simplest form for this is to include another binary predictor variable, `resit`, which is `1` for resit marks and `0` otherwise. In the projects where I have applied these models, I have excluded the rare case of third attempts.

```
// priors:
beta_0 ~ uniform(-30, 30)
beta_1 ~ uniform(-30, 30)
beta_2 ~ uniform(-30, 30)
```

*especially in the version of MCMC used in BUGS and JAGS (the Gibbs sampler), because suddenly every other parameter's values are all extremely unlikely in all directions and it is hard to find the high-posterior values that are now concentrated in a very small region

```
beta_3 ~ uniform(-30, 30)
sigma ~ uniform(0.01, 25)
tau_u ~ uniform(0.01, 25)
tau_v ~ uniform(0.01, 25)
for(i in 1:n_students) {
   u[i] ~ normal(0, tau_u)
}
for(j in 1:n_courses) {
   v[j] ~ normal(0, tau_v)
}

// likelihood:
for(k in 1:n_marks) {
   yhat[k] = beta_0 + beta_1*socio_economic[k] +
             beta_2*intervention[k] +
             beta_3*resit[k] +
             u[student_id[k]] -
             v[course_id[k]]
   y[k] ~ normal(yhat[k], sigma)
}
```

We have given `beta_3` a uniform prior centred on zero, just like the other coefficients, but we would probably be justified to push the resit effect towards being positive, because that is every educator's experience (averaging across students) and because there will not be many resits and the prior will be more influential than for the other coefficients.

In many institutions, a student who fails, then passes at resit will have their marks "capped" at the pass mark. If anything at or above 40% is a pass, for example, then all the resit pass marks will be recorded as 40%. We will tackle this in the next subsection.

3.3 Coarse data

Resits capped at the pass mark only tell us that the student obtained a mark at or above that pass mark. In other words, they obtained a mark between 39.5% (because it is rounded to the nearest percentage) and 100%. Similarly, some student databases might record all failed assessments as 0%, which tells us that they

obtained something between 0% and 39.5%.

Similar problems arise when some courses or institutions only record a grade (for example, A+, A, B, and so on), or a binary pass/fail (such as in clinical placements for health professions). Students who do not attend an examination or submit an assignment without acceptable mitigating circumstances might be automatically given 0%, but if they had undertaken the assessment, they would probably not have obtained exactly 0%.

There is a generic framework to modelling *coarse* data like these.[6] Because we always know which marks are coarsened and which are not, it is int his case equivalent to what biostatisticians might call interval censoring. Missing data, often incorporated nowadays via a procedure called multiple imputation, is just a special case of coarse data; if we had truly missing marks because of a database problem, we would only know that the true value lay between 0% and 100%.

Instead of having a precise value for the coarse marks, Bayesian modelling allows us to give these students a probability distribution. Their "true" attainment is another unknown, and we can allocate a prior to it. Less satisfactory approaches might be to exclude all such data (thus increasing uncertainty in the analysis, and possibly biasing it) or to replace each of these coarse marks with a single imputed value, such as the average of the pass marks at first sitting.

There are two possible approaches for allocating priors.

In the first, we believe it justified to model all the coarsened marks as having the same prior distribution. Consider the simplest case, where database failure has wiped out some students' marks, completely at random. We might give them the distribution that best approximates the marks that we do know about, on the basis that they should be just like the rest of the complete data. This is called Coarsened Completely At Random (CCAR).

Those who are capped at resit could be given the part of this probability distribution above 39.5% (and it will have to be rescaled so that it integrates to one, as all probability distributions should). Those coarsened to 0% get the part below 39.5%, also rescaled. All we are asking the computer to do is to predict what these marks might have been via posterior distributions. The posteriors will still be influenced by the other independent variables.

The second approach is to give them all the same distribution, but also to predict a binary indicator variable for each type of coarse value; `coarse40` would be `1` if they are coarsened at 40% and `0` otherwise, and there might be `coarse0` and `missing` as well. We predict each of these with another model, the *coarsening*

model, for which we do not have to be constrained to the variables used in our *substantive* model: `socio_economic`, `intervention`, and `resit`. Anything that might predict their performance could be considered.

This approach is called Coarsened At Random (CAR); it effectively asserts that, once we know all the predictor variables for coarsening, we don't need any other information to be able to predict it. Crucially, we do not need to know the missing mark itself. Because these new variables are binary, we use a Bernoulli distribution to give us the probability of them being `1`.

3.3.1 Coarsened Completely At Random: imputation

Let's start with CCAR. The code below fits the substantive model and predicts coarse data, where there is only one type of coarseness: some fails are recorded as 0%, for no particular reason, and we know exactly which those are (as opposed to any students who might truly have got 0%), so there is no coarsening model[†], and we merely impute the true attainment using the substantive model.

We must supply the data in two parts: `n_complete` rows with complete marks and `n_0` rows with coarsened zeros. We will use the `normal_truncated()` probability distribution for the imputed marks, as they follow a normal distribution according to the substantive model, but must be below 39.5%. There are three arguments: the mean, standard deviation, lower and upper limits. The lower limit is omitted, so it is not constrained below.

```
// priors:
beta_0 ~ uniform(-30, 30)
beta_1 ~ uniform(-30, 30)
beta_2 ~ uniform(-30, 30)
beta_3 ~ uniform(-30, 30)
sigma ~ uniform(0.01, 25)
tau_u ~ uniform(0.01, 25)
tau_v ~ uniform(0.01, 25)
for(i in 1:n_students) {
   u[i] ~ normal(0, tau_u)
}
```

[†]This is, of course, not very realistic. Usually, there is very valuable contextual information that helps us define a better model, and that information is obtained by talking to people involved in managing the student database, the instructors, and so on.

```
for(j in 1:n_courses) {
   v[j] ~ normal(0, tau_v)
}

// likelihood, substantive model:
for(k in 1:n_complete) {
   yhat[k] = beta_0 + beta_1*socio_economic[k] +
             beta_2*intervention[k] +
             beta_3*resit[k] +
             u[student_id[k]] -
             v[course_id[k]]
   y[k] ~ normal(yhat[k], sigma)
}

// imputation:
for(k in 1:n_0) {
   yhat_imputed[k] = beta_0 + beta_1*socio_economic[k] +
                     beta_2*intervention[k] +
                     beta_3*resit[k] +
                     u[student_id[k]] -
                     v[course_id[k]]
   y_imputed[k] ~ normal_truncated(yhat_imputed[k],
                                   sigma,
                                   ,39.5)
}
```

You might very well object to the normal distribution here, as we know there are also constraints at 0% and 100%. That is quite reasonable, and an alternative might be to use a truncated normal for the substantive model too, or to divide percentages by 100, so the marks are between 0 and 1, then use beta distributions instead.

I will not go into this further here, other than to say that it is a good idea, but rarely if ever seen in practice, and therefore likely to attract undue scepticism from reviewers and readers who feel that you may be hiding something or that your paper/report is too methodological. You will have to explain the methods more, and so it will increase word count. If the normal distribution has little area under the curve outside [0%, 100%], then it will make little difference anyway. However,

implementing it in a PPL is as easy as swapping the `normal()` distribution function for a `beta()`. Because the beta's parameters are not a mean and standard deviation, you would then need to calculate those in additional deterministic lines, but that is quite simple once you look up the formula.

3.3.2 Coarsened Completely At Random: Bernoulli and normal likelihoods

Now, I will extend what we have just done to a setting where some assessments are only recorded as pass or fail. It might be that there was a percentage mark, but it is not recorded, in which case we want to include the coarse data rather than deleting it from the analysis, or some crude single imputation such as replacing all fails with exactly 30% and all passes with exactly 60%.

Supposing that they were only ever marked as pass/fail, we can say that the assessments have a true, *latent attainment*, and we might expect those to be distributed rather like the percentages that we know in full. We can cut those latent marks at 39.5% and allocate those below to fail and those above to pass. This latter way of thinking about it actually gives us an inroads to connecting the fully known, normally distributed, percentages with the binary pass/fail.

We use a `bernoulli()` distribution for binary data (`0` for fail, `1` for pass). It takes only one argument, the probability of being `1`. And if we measure the area under the curve of the normal distribution of the latent marks, we can cut this into the probability of being anywhere below 39.5%, or above (Figure1). The area under the curve up to a point is called the cumulative distribution function, or CDF. Mathematically minded readers might like to note that it is the integral of the probability density function, or PDF. To get the probability of being *above* 39.5%, we just subtract it from 1.

In our PPL, we will follow these steps:

- calculate the likelihood of the fully-known marks from the normal likelihood, as before
- calculate this CDF, contingent on the substantive model parameters, using a built-in function `normal_cdf()`
- subtract it from 1 to get the probability of passing
- plug the probability of passing into the Bernoulli likelihood for the pass/fail marks

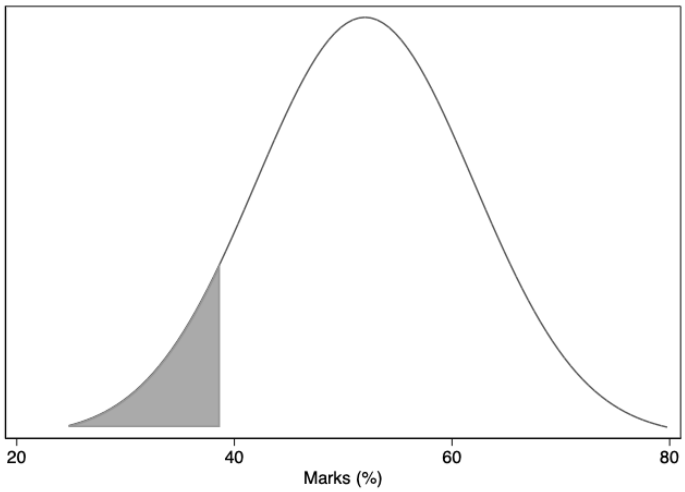

Figure 1: Normal PDF with mean 52% and standard deviation 10%. The CDF below 39.5% is shaded in grey and amounts to 0.1056, or a 10.56% probability of failing.

The PPL takes care of the MCMC process behind the scenes, using these likelihoods and priors to sample from the posterior.

```
// priors:
beta_0 ~ uniform(-30, 30)
beta_1 ~ uniform(-30, 30)
beta_2 ~ uniform(-30, 30)
beta_3 ~ uniform(-30, 30)
sigma ~ uniform(0.01, 25)
tau_u ~ uniform(0.01, 25)
tau_v ~ uniform(0.01, 25)
for(i in 1:n_students) {
   u[i] ~ normal(0, tau_u)
}
for(j in 1:n_courses) {
   v[j] ~ normal(0, tau_v)
}

// likelihood, substantive model with complete data:
for(k in 1:n_complete) {
```

```
    yhat[k] = beta_0 + beta_1*socio_economic[k] +
              beta_2*intervention[k] +
              beta_3*resit[k] +
              u[student_id[k]] -
              v[course_id[k]]
    y[k] ~ normal(yhat[k], sigma)
}

// likelihood, binary data:
for(k in 1:n_binary) {
    yhat_imputed[k] = beta_0 + beta_1*socio_economic[k] +
                      beta_2*intervention[k] +
                      beta_3*resit[k] +
                      u[student_id[k]] -
                      v[course_id[k]]
    prob_fail = normal_cdf(yhat_imputed[k],
                           sigma,
                           39.5)
    prob_pass = 1-prob_fail
    y_binary[k] ~ bernoulli(prob_fail)
}
```

3.3.3 Coarsened At Random: coarsening model and substantive model

Although it seems unlikely in an education setting, it could conceivably be the case that the coarsening is not completely at random: not done for all assessments in a certain course, or for a random selection of assessments, but instead because of other factors *that we know about*. Imagine a project that looks at student marks from a university database, going back over ten years. In the earlier part of this, the data entry system was different, and only binary pass/fail was recorded. But, problematically, it is also known that attainment has improved at the university over this time period (or exams have gotten easier!), so we cannot use the fully-known (later years) marks to calculate a substantive model and hence the probability of fails and passes in the binary (early years) data. We must take the year into account.

It is important to note that ignoring the year will bias the results because year both caused the coarsening and influenced the marks. This makes it a form of confounder, or as Pearl's approach to causal inference would explain it, a backdoor

path from the coarseness to the marks.[7] Not every factor that explains coarsening is associated with the mark, and those that are not need not be used in the model as they will not in general induce bias.

This calls for a model that imputes the latent marks, and hence the Bernoulli likelihood, using substantive predictors and year. Year does not have to be in the substantive model to do this. It might only differ from the previous PPL code by including another variable and coefficient in the deterministic line calculating `yhat_imputed[k]`.

3.3.4 Generalized coarsening

Most of Heitjan and Rubin's work on coarsening[6] focusses on a situation where we might not know whether a mark has been coarsened or not. This is unlikely in education, but I will outline it.

They present an example in another paper using smoking statistics; smokers are asked how many cigarettes they smoke per day. There is an excess of numbers rounded to multiples of 5 and 10, a data artefact called heaping. The researchers suspect that this is just a mental rounding-off, and in fact likely to be rounding-down. They use a coarsening model to predict the binary status of being coarsened or not, and this can contain any variables that improve the model fit: its goal is prediction, not causal inference. This brings such models very much in line with missing data models, including those that use multiple imputation.

3.3.5 Coarsening Not At Random: problematic data, sensitivity analysis

The categories CCAR and CAR are based on Rubin's earlier work on missing data. He defined a third category, Missing Not At Random, where the probability of being missing was determined by factors that influenced the outcome, but were *not* contained in the data at hand. The equivalent Coarsening Not At Random (CNAR) would mean that any analysis is liable to be biased because the coarse data are not like the fully known data.

If we set up a Bernoulli likelihood for the coarse data which has entirely different coefficients to the normal likelihood for the fully-known data, we will effectively be running a logistic regression on the coarse data and a linear regression on the fully-known. Because there are no shared data and no shared parameters, we are unable to take advantage of both sets of data and allow them to "pool information".

The only universal solution is to find out (by debriefing, ethnographic or other qualitative research) the guilty factors, then collect a new dataset including them. However, in Bayesian statistics we can take advantage of the ability to use priors for any unknown, and set up a prior distribution for the size of the difference between the means in the fully-known marks and the latent marks. This could be defined by a consensus elicitation process.[8] In some ways, this is preferable for its concise, single-parameter conclusion about coarsening, allowing simpler communication and validation. On the other hand, it puts the conclusions at the mercy of unquestioned assumptions, prejudices and *ipse dixit*, because there are no data to allow checking.

4 Output: communicating results

Bayesian methods give us a number of different options for communicating findings, compared to frequentist* methods. We do not obtain p-values, and therefore we cannot say if a difference or trend is "significant". Collaborators can find this troubling, as the p-value is the great signifier of quantitative science in our time. However, there are excellent alternatives that become possible, and I would argue superior alternatives.

We obtain an estimate of the posterior probability distribution for the parameters, and we can derive various useful measures from that. Continuing the hypothetical example from above, if we are principally interested in the impact of our intervention on the marks, we should describe our best guess at the value of `beta_2`, and also quantify the uncertainty around that best guess.

Usually, we use methods like MCMC, which do not give us a formula for the posterior, but a sample containing many draws from that posterior distribution. To summarise the posterior, we can just calculate summary statistics of the sample. The "best guess" could be the mean of the sample, although some prefer to estimate the mode – the parameter value where the posterior probability is highest. The mode requires some further computation while the mean is simply all the draws in the sample added together and divided by the number of draws in the sample.

The uncertainty around the best guess is where Bayesian inference gets interesting, because we have incorporated different types and sources of uncertainty

*Many of the techniques that are commonly lumped under this term do not really rely on a specific interpretation of probability as a frequency of similar findings from a long run of identical data collections. I prefer to say Bayesian and non-Bayesian, and to divide up statistical methods on other aspects,[9] but I will use the familiar term here.

in our model and now we can see their combined effect. One simple approach is to give a 95% credible interval, which is the interval in which the true effect of the intervention is 95% likely to lie. This is a much simpler interpretation than the traditional frequentist one, and closer to what most of our audience is thinking anyway.

The simplest way to get the 95% credible interval is just to find the 2.5 centile and the 97.5 centile of the posterior sample. Of course, there is nothing magic about 95% and other intervals can be derived instead of, or alongside it. We might even visualise the distribution in a histogram or kernel density plot.

We can evaluate the area under the curve above zero: the probability that the intervention is beneficial. But some of that curve could be beneficial by a very small amount. The Bayesian interpretation gives further value when we can compare posterior probabilities to some contextual value. For example we could state the probability that our new intervention will have a bigger impact than some old intervention. Frequentist outputs cannot tell you this.

In medical research, it is common to convene a group of experts before a study begins, to decide on what would constitute a "minimal clinically important difference". This is a value for the effect of the intervention where we shuld start to be interested in making recommendations to use it. If an intervention increases student marks by 0.1% on average, who cares? It might even be statistically significant, in that it is unlikely to be an artefact of a truly zero effect, but it is not of interest in practical terms. Perhaps we decide that only interventions achieving a 5% increase on average are of interest. Now, we can provide the posterior probability that the new intervention will achieve (or exceed) that level. Again, frequentist outputs cannot tell you this.

There are occasions when a truly binary decision must be made. A university either implements the new intervention or does not. Frequentist significance appears to guide such decisions, but often the decision does not end with whether the data could have arisen from a truly zero effect. Instead, the decision-makers must consider how likely it is that the effect could be over some threshold. They might, before collecting the data, decide on an acceptable risk: for example, if the new intervention has a 60% or higher probability of delivering a 5% or more improvement in marks. They could also act negatively, to rule out risk of wasting money or harming students: to implement it unless there is more than a 30% chance that it will on average reduce student marks. These are all just different ways of dividing up the posterior probability and assessing the area under the curve.

How do we assess this area under the curve when what we have is a sample of a few thousand draws from the posterior distribution? We simply count how many are above or below some threshold.

In Section 7, I describe a case study where colleagues and I chose a pre-determined threshold for importance of a parameter, and defined a substitute for significance on that basis.

We can respond to the need to evaluate how well our models fit the data, even to compare multiple models, by using prior and posterior predictive checking. The prior predictive approach generates new pseudo-data according to the prior probabilities and the model. If the results do not encompass the observed data, then the priors need to be adjusted as they will never be able to fit the data. Posterior predictive checking uses the posterior distributions instead, and visually, we can consider whether there are patterns evident in the data that are not captured by the model. If so, we should adjust the model.

Some Bayesian analysts respond to the need for hypothesis testing by using Bayes factors, a measure of how (posterior) probable hypothesis A is compared to hypothesis B. There are even some widely-used thresholds for the Bayes factor to help guide binary decisions.[10]

The fact that we have a posterior sample allows us to easily use that information on uncertainty for other purposes; it is not an obscure formula but can be presented as a collection of potential realities, one of which is true, but we don't know which. Posterior samples can be supplied as an input to economic evaluation, and the multiple potential realities can be shown superimposed in data visualisation.[11,12]

The estimates of student abilities and course difficulties can be used (and were used, in the case studies below) to validate the analysis. Although the statistician should have anonymised data, a collaborator involved in the education programme under study will have a key file, which relates each student's real university ID to the meaningless study ID. This collaborator could be given a list of the student IDs associated with the top 10 and bottom 10 student abilities, and likewise for course difficulties, which they could discreetly check against their own list and from their experience of the students, confirm if this makes sense or if something does not seem to match reality.

If required, courses could be ranked in terms of difficulty visually, taking into account the uncertainty from the model and the data by means of graphics used in performance indicators, including caterpillar plots and funnel plots.[13] Early detection of variation between courses, departments, institutions or instructors can be

informed by cumulative funnel plots,[14] or similar statistical process control charts.

5 Software

BUGS[1] and JAGS[5] are both implementations of an efficient version of MCMC called the Gibbs sampler. BUGS can be used from within R, Python, Stata or standalone as the package WinBUGS (Windows only). JAGS can be used from within R, Python, Stata, Julia (and probably other packages too).

Stan[2] is an alternative that uses a much improved version of MCMC called Hamiltonian Monte Carlo. This is more stable and much faster for multilevel models such as those set out in this paper. It can be called from within R, Python, Stata, Julia, Matlab, and others, or in a standalone command-line program called CmdStan.

There are some higher-level packages such as brms in R, and PyMC in Python, which do not require a full PPL specification of the model. However, I strongly recommend the PPl approach, because it maps more directly to how non-experts in statistics might conceive of a model. This makes it easier to discuss and reach informed consensus on model specification and also to spot coding errors.

There are many other Bayesian sampling algorithms, which the reader may see named from time to time, such as INLA (integrated nested Laplace approximations), variational inference, particle filters, ABC(approximate Bayesian computation), MALA (Metropolis-adjusted Langevin algorithm) and PDMPs (piecewise deterministic Markov processes) such as the zig-zag sampler or bouncy particle sampler. in practice, for models such as are described in this paper, they represent no benefit in time and stability, and introduce risks in their less familiar behaviour. They have important roles to play, but this is not one.

Many educational researchers might be using older software such as SPSS. Unfortunately, there is no way at present to integrate effective Bayesian sampling with a PPL into such software.

6 Case study 1: multiple mini-interviews for nursing students

This project examined nursing students at one university.[15] All applicants had completed a process called multiple mini-interviews (MMI), which provided scores for numeracy, communication skills and an overall score. The goals were to consider

whether there may have been biases in how the MMI system handled applicants from different backgrounds, and whether MMI scores could predict marks, as the MMI developers had hoped.

The data required an IRT model with effects of student ability, course ("module", in this setting) difficulty, and resits ("attempts"). Those recorded only as fail or given 0% by course regulations were imputed CCAR to the truncated predicted distribution below 39/5%, and those recorded only as pass or capped at 40% were similarly imputed above 39.5%. A pseudo-code version of the model follows:

```
// priors:
sigma ~ normal_truncated(0, 30, 0, )
alpha ~ normal(50, 30)
beta ~ normal(0, 20)
sigma_stu ~ normal(0, 30, 0, )
sigma_mod ~ normal(0, 30, 0, )

// priors for random effects:
for (i in 1:N_students) {
   ustu[i] ~ normal(0, sigma_stu)
}
for (i in 1:N_modules) {
   umod[i] ~ normal(0, sigma_mod)
}

// likelihood for complete data:
for (i in 1:Nmark) {
   mu[i] = alpha+(beta[1]*mmi_interview[i])+
                 (beta[2]*mmi_maths[i])+
                 (beta[3]*mmi_composition[i])+
                 (beta[4]*attempt[i])+
                 ustu[student_id[i]]+
                 umod[module_id[i]]
   mark[i] ~ normal(mu[i], sigma)
}
for (i in 1:N0) {
   mu0[i] = alpha+(beta[1]*mmi_interview0[i])+
                  (beta[2]*mmi_maths0[i])+
```

```
                    (beta[3]*mmi_composition0[i])+
                    (beta[4]*attempt0[i])+
                    ustu[student_id0[i]]+
                    umod[module_id0[i]]
    latentmark0[i] ~ normal(mu0[i], sigma)
}
for (i in 1:N40) {
    mu40[i] = alpha+(beta[1]*mmi_interview40[i])+
                    (beta[2]*mmi_maths40[i])+
                    (beta[3]*mmi_composition40[i])+
                    (beta[4]*attempt40[i])+
                    ustu[student_id40[i]]+
                    umod[module_id40[i]]
    latentmark40[i] ~ normal(mu40[i], sigma)
}
```

In this project, we chose only to impute the binary marks for two reasons: first, those modules were in the minority and their inclusion would make little impact on the likelihood, and secondly, as clinical placements they were completely different to the other, academic, modules, and the distribution of their latent marks was probably unlike that of the academic marks. Clinical placements, my colleagues advised, were designed to assess completely different competencies.

7 Case study 2: widening participation in physiotherapy students

This project was similar in data structure to the MMI project, but included data from more than one institution. At the outset, to obtain consent from as many institutions as possible, collaborators agreed that the analyses would not evaluate or compare the institutions. So, we did not include university as a third layer in the model. Each university contributed its own courses, and so we expect all the course-specific effects, taken together, to follow one normal distribution.

The goal of the project was to assess whether students taking physiotherapy degrees (there were separate analyses, of the same model, for BSc and MSc data) achieved broadly the same marks, or whether there were differences along: gender, age, ethnicity, disability, home postcode socio-economic status, and previous quali-

fications. We also considered whether different types of assessment — blinded and untimed (e.g. written assignments), blinded and timed (e.g. exminations), or unblinded (e.g. objective structured clinical examinations (OSCEs): clinical role-play scenarios) — impacted differently on the various student groups, using interaction terms in a linear regression (the `gamma` matrix in the pseudo-code below).

```
sigma ~ normal_truncated(0, 30, 0, )
alpha ~ normal(50,30)
beta ~ normal(0,20);
sigmastu ~ normal(0,30);
ustu ~ normal(0,sigmastu);
sigmamod ~ normal(0,30);
umod ~ normal(0,sigmamod);
for (i in 1:Nmark) {
   mu[i] = alpha+
          (beta[1]*blinded_untimed1[i])+
          (beta[2]*blinded_timed1[i])+
          (beta[3]*clinical1[i])+
          (beta[4]*female1[i])+
          (beta[5]*over201[i])+
          (beta[6]*ethn_asian1[i])+
          (beta[7]*ethn_black1[i])+
          (beta[8]*ethn_other1[i])+
          (beta[9]*disab_phys1[i])+
          (beta[10]*disab_learn1[i])+
          (beta[11]*disab_other1[i])+
          (beta[12]*polar1[i,1])+
          (beta[13]*polar1[i,2])+
          (beta[14]*polar1[i,3])+
          (beta[15]*polar1[i,4])+
          (beta[16]*qual_access1[i])+
          (beta[17]*qual_degree1[i])+
          (beta[18]*qual_other1[i])+
          (gamma[1,1]*ethn_asian1[i]*blinded_untimed1[i])+
          (gamma[2,1]*ethn_black1[i]*blinded_untimed1[i])+
          (gamma[3,1]*ethn_other1[i]*blinded_untimed1[i])+
          (gamma[1,2]*ethn_asian1[i]*blinded_timed1[i])+
```

```
          (gamma[2,2]*ethn_black1[i]*blinded_timed1[i])+
          (gamma[3,2]*ethn_other1[i]*blinded_timed1[i])+
          (gamma[1,3]*ethn_asian1[i]*clinical1[i])+
          (gamma[2,3]*ethn_black1[i]*clinical1[i])+
          (gamma[3,3]*ethn_other1[i]*clinical1[i])+
          ustu[stu1[i]]+umod[mod1[i]]
   mark[i] ~ normal(mu[i], sigma)
}
for (i in 1:N0) {
   mu0[i] = alpha+
          (beta[1]*blinded_untimed0[i])+
          (beta[2]*blinded_timed0[i])+
          (beta[3]*clinical0[i])+
          (beta[4]*female0[i])+
          (beta[5]*over200[i])+
          (beta[6]*ethn_asian0[i])+
          (beta[7]*ethn_black0[i])+
          (beta[8]*ethn_other0[i])+
          (beta[9]*disab_phys0[i])+
          (beta[10]*disab_learn0[i])+
          (beta[11]*disab_other0[i])+
          (beta[12]*polar0[i,1])+
          (beta[13]*polar0[i,2])+
          (beta[14]*polar0[i,3])+
          (beta[15]*polar0[i,4])+
          (beta[16]*qual_access0[i])+
          (beta[17]*qual_degree0[i])+
          (beta[18]*qual_other0[i])+
          (gamma[1,1]*ethn_asian0[i]*blinded_untimed0[i])+
          (gamma[2,1]*ethn_black0[i]*blinded_untimed0[i])+
          (gamma[3,1]*ethn_other0[i]*blinded_untimed0[i])+
          (gamma[1,2]*ethn_asian0[i]*blinded_timed0[i])+
          (gamma[2,2]*ethn_black0[i]*blinded_timed0[i])+
          (gamma[3,2]*ethn_other0[i]*blinded_timed0[i])+
          (gamma[1,3]*ethn_asian0[i]*clinical0[i])+
          (gamma[2,3]*ethn_black0[i]*clinical0[i])+
```

```
          (gamma[3,3]*ethn_other0[i]*clinical0[i])+
          ustu[stu0[i]]+umod[mod0[i]]
latentmark0[i] ~ normal(mu0[i], sigma)
}
for (i in 1:N40) {
   mu40[i] = alpha+
          (beta[1]*blinded_untimed40[i])+
          (beta[2]*blinded_timed40[i])+
          (beta[3]*clinical40[i])+
          (beta[4]*female40[i])+
          (beta[5]*over2040[i])+
          (beta[6]*ethn_asian40[i])+
          (beta[7]*ethn_black40[i])+
          (beta[8]*ethn_other40[i])+
          (beta[9]*disab_phys40[i])+
          (beta[10]*disab_learn40[i])+
          (beta[11]*disab_other40[i])+
          (beta[12]*polar40[i,1])+
          (beta[13]*polar40[i,2])+
          (beta[14]*polar40[i,3])+
          (beta[15]*polar40[i,4])+
          (beta[16]*qual_access40[i])+
          (beta[17]*qual_degree40[i])+
          (beta[18]*qual_other40[i])+
          (gamma[1,1]*ethn_asian40[i]*blinded_untimed40[i])+
          (gamma[2,1]*ethn_black40[i]*blinded_untimed40[i])+
          (gamma[3,1]*ethn_other40[i]*blinded_untimed40[i])+
          (gamma[1,2]*ethn_asian40[i]*blinded_timed40[i])+
          (gamma[2,2]*ethn_black40[i]*blinded_timed40[i])+
          (gamma[3,2]*ethn_other40[i]*blinded_timed40[i])+
          (gamma[1,3]*ethn_asian40[i]*clinical40[i])+
          (gamma[2,3]*ethn_black40[i]*clinical40[i])+
          (gamma[3,3]*ethn_other40[i]*clinical40[i])+
          ustu[stu40[i]]+umod[mod40[i]]
   latentmark40[i] ~ normal(mu40[i], sigma)
}
```

8 Case study 3: assessing the impact of training on palliative care professionals

While the two previous models had only one outcome variable, marks, and assumed a normal distribution for those, hence linear regression, the third case study involved ordinal data on several different questions about confidence.

This collaboration with Princess Alice Hospice evaluated participants in the European Certificate in Essential Palliative Care, before the course, at the end of the course, and 3 months later. There was therefore a time coefficient, of principal interest in this study. Students each had their own student ability random effect, thus naturally allowing for missing responses at the various time points.

There were questions on various topics, such as psychological support or pain relief, and these topics recurred (though not completely) across management, multidisciplinary working, and communication. Therefore, there were also topic difficulty random effects. Some communication topics were repeated with relation to the patient and the patient's family. We included an additional parameter, `pt`, to capture the relative confidence or lack of confidence in talking to the patient themselves. The participant's sex, time in post, work setting and profession were included as other predictors. The goal was to estimate the impact of the training on confidence topics, and whether any benefit was sustained after 3 months. There was no imputation of missing responses.

In the pseudo-code below, a latent confidence value for each student and topic is subdivided into nine ordinal categories. We used an ordinal logistic model (see Figure 2), which turns a latent confidence distribution into a series of proportions by cutting it at several points (determined by the computer). Such models are available as presets in BUGS, JAGS and Stan; here I call it `ordinal_logistic()`.

This code is more compact than the previous case studies, in part because it makes more use of vectors and matrices to contain the parameters and data. Readers unfamiliar with matrix multiplication can still get the general idea by thinking about just one element of vectors `beta`, `beta_p` and `c`, and one row of data from matrices `x` and `xp`.

```
beta ~ normal_vector(0, 20, n_predictors)
beta_p ~ normal_vector(0, 20, n_baseline_predictors)
pt ~ uniform(-50, 50)
sigma_stu ~ normal(0,20)
sigma_topic ~ normal(0,20)
```

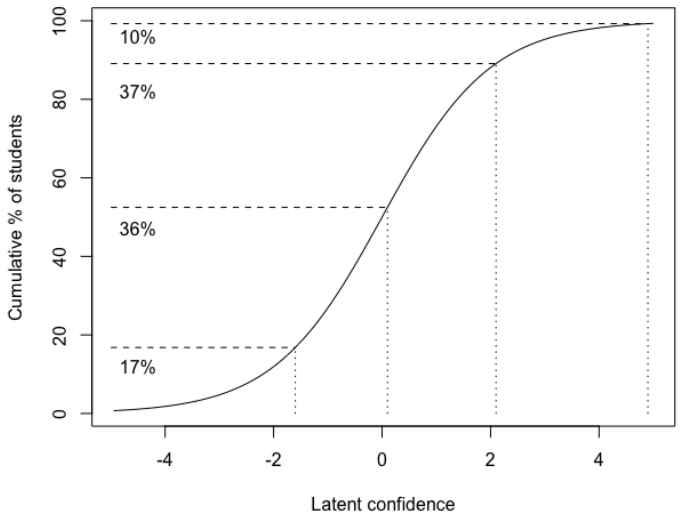

Figure 2: Logistic CDF of the latent confidence, cut at the vertical lines into four ordinal categories. These are converted to proportions (which add to 100%) seen on the left.

```
for (i in 1:N_students) {
   u_stu ~ normal(0, sigma_stu)
}
for (i in 1:N_topics) {
   u_topic ~ normal(0,tau_topic)
}
// prior for 8 cutpoints:
c ~ normal_ordered_vector(0, 20, 8)
for (i in 1:N_students) {
   // each participant's baseline confidence
   // xp is a matrix of participant-level
   //    predictors, e.g. profession
   mu[i] = u_stu[i] + xp[i]*beta_p
}
for (i in 1:N) {
     d_effects_pt[i] ~ ordered_logistic(u_topic[1]+pt+
                       mu[stu_id[i]]+(x[i]*beta), c)
     d_effects_fam[i] ~ ordered_logistic(u_topic[1]+
```

```
                    mu[stu_id[i]]+(x[i]*beta), c)
  d_dying[i] ~ ordered_logistic(u_topic[2]+
                    mu[stu_id[i]]+(x[i]*beta), c)
  d_death_pt[i] ~ ordered_logistic(u_topic[3]+pt+
                    mu[stu_id[i]]+(x[i]*beta), c)
  d_death_fam[i] ~ ordered_logistic(u_topic[3]+
                    mu[stu_id[i]]+(x[i]*beta), c)
  d_bereavement[i] ~ ordered_logistic(u_topic[4]+
                    mu[stu_id[i]]+(x[i]*beta), c)
  d_prognosis[i] ~ ordered_logistic(u_topic[5]+
                    mu[stu_id[i]]+(x[i]*beta), c)
  d_pain[i] ~ ordered_logistic(u_topic[6]+
                    mu[stu_id[i]]+(x[i]*beta), c)
  assess[i] ~ ordered_logistic(u_topic[7]+
                    mu[stu_id[i]]+(x[i]*beta), c)
  aetiology[i] ~ ordered_logistic(u_topic[8]+
                    mu[stu_id[i]]+(x[i]*beta), c)
  symptoms[i] ~ ordered_logistic(u_topic[9]+
                    mu[stu_id[i]]+(x[i]*beta), c)
  prescribe[i] ~ ordered_logistic(u_topic[10]+
                    mu[stu_id[i]]+(x[i]*beta), c)
  effects[i] ~ ordered_logistic(u_topic[1]+
                    mu[stu_id[i]]+(x[i]*beta), c)
  psychological[i] ~ ordered_logistic(u_topic[11]+
                    mu[stu_id[i]]+(x[i]*beta), c)
  social[i] ~ ordered_logistic(u_topic[12]+
                    mu[stu_id[i]]+(x[i]*beta), c)
  spiritual[i] ~ ordered_logistic(u_topic[13]+
                    mu[stu_id[i]]+(x[i]*beta), c)
  mdt[i] ~ ordered_logistic(u_topic[14]+
                    mu[stu_id[i]]+(x[i]*beta), c)
  ref_physio[i] ~ ordered_logistic(u_topic[15]+
                    mu[stu_id[i]]+(x[i]*beta), c)
  ref_ot[i] ~ ordered_logistic(u_topic[16]+
                    mu[stu_id[i]]+(x[i]*beta), c)
  ref_complementary[i] ~ ordered_logistic(u_topic[17]+
```

```
                    mu[stu_id[i]]+(x[i]*beta), c)
    ref_lymphoedema[i] ~ ordered_logistic(u_topic[18]+
                    mu[stu_id[i]]+(x[i]*beta), c)
    ref_psychiatric[i] ~ ordered_logistic(u_topic[11]+
                    mu[stu_id[i]]+(x[i]*beta), c)
    ref_spiritual[i] ~ ordered_logistic(u_topic[13]+
                    mu[stu_id[i]]+(x[i]*beta), c)
}
```

The deterministic aspects of the model happen in two steps. First, the participant-level predictors`xp` are combined with their coefficients `beta_p` to predict `mu`, a vector with a value for each participant. But participants have data at up to three time points, so, for each of the 23 questions, `mu` is combined with the topic difficulty `u_topic`, indicator variables for the time points in the matrix `x` and the patient effect `pt` to obtain a value that is an input to `ordered_logistic`.

Because the inverse logistic function — the ramp shape in Figure 2 — has a fixed midpoint (zero) and spread, the `ordered_logistic()` function takes two inputs: a value, predicted by the model formulas, and a vector of cutpoints. It returns a code for one of the categories; if there are 8 cutpoints, they define 9 categories. As the posterior distribution produces a range of different input predictions, so those are mapped to the categories.

Typically, we are interested in whether a predictor, such as time in this case, affects results (seen as a posterior distribution for relevant components of `beta` that is above zero to an extent defined before the study began). However, we might also want to compare the predicted and observed confidence levels. For this, we can store posterior predictive outputs from the `ordered_logistic()` functions, and simply describe the frequency with which various categories occur in the posterior sample, which will be the posterior probability of a particular participant, at a particular time point, responding in a particular category.

9 Practicalities

Each of the three case studies was built up, step by step, from extremely simple linear regression models, so that problems arising at any step could be isolated and corrected. They were programmed in Stan, and the original code is available

online.[18] Fitting the full models in Stan with thousands of observations will take a few hours on a good consumer-grade computer, but it is possible to use parallel computation to speed this up further, especially with newer versions of CmdStan. A affordable cloud computing facility such as DigitalOcean can be used for this. Anyone is welcome to adapt the pseudo-code here, as it is published under The Unlicense, but you should exercise great caution in adapting it to your own needs and software.

It is important for organisations and management to have a clear and realistic standard in mind for the human expertise required as an input to such a project. While statistician input is at a premium in many research settings, particularly outside science faculties, it is essential to have some level of expert oversight. The models described here lend themselves to PPL software, all of which are open-source and represent no extra expense to the researcher. The expense, however, is in the form of recruitment and training.

Statisticians interested in Bayesian methods should find this an attractive collaboration prospect, as it is relatively *prêt à analyser*, while retaining some scope for personalisation, and hence spin-off methodological outputs. It is important to involve them from the outset, before data are even assembled, and to remember that an academic statistician operating at this level of sophistication is not a helpdesk, and may choose not to collaborate at all on such terms.

Qualifications involving machine learning and data science are increasingly widespread, and it may be tempting to allocate a junior member of staff from that background to provide the analysis. Managers may not be aware of the distinction, but in a nutshell, machine learning methods provide estimates without uncertainty, while the entire focus of Bayesian statistics in on understanding uncertainty. Also, there are some machine learning methods which involve the name of the Reverend Thomas Bayes, but are quite unrelated, for example the "naive Bayes classifier". Unfortunately, the manager seeking collaboration, or recruiting, has to be able to bear these pitfalls in mind.

Another option is to bring in a statistical consultancy, although there are few specialising in Bayesian modelling expertise, and with a proven track record of working alongside non-statistical researchers to define a project and bring it to fruition, including effective communication of findings.

10 Au-delà

There are many ways in which these models can be expanded. Perhaps the most relevant for educational researchers is to incorporate structural equation modelling, or confirmatory factor analysis.[19] When we know *a priori* that certain topics are linked, we can constrain the connections between variables so that, for example, those topics share a common input.

Students' previous experience of caring for a family member, for example, may have a positive effect on attainment in clinical placement, but not in courses about physiology or research methods. Considering post-employment follow-up data, we might expect clinical placement and physiology performance to impact confidence in providing care, and research methods to impact confidence in contributing to multi-disciplinary discussions. These links can be built into the model code. Competing models of structure can be compared objectively using measures such as information criteria, Bayes factors or more open-ended investigations such as posterior predictive checking.

We explored imputation of coarse data CAR or CNAR above, and the same idea applies to missing data,[20] which is often a problem not in the assessment outcomes, but in the predictor variables that we might wish to collect through student surveys or pre-admission questionnaires. In a Bayesian framework, missing data are simply more parameters, with prior distributions informed by the known data.

11 Conclusion

This form of model for educational assessment data is practicable, given a certain level of statistical expertise. There is no need for more than free, open-source software. It provides several advantages over simplifications that might squeeze complex and contextually nuanced data into a simple hypothesis-testing framework. I encourage all education researchers to investigate it further and help to raise the standard of these analyses.

References

[1] Lunn D, Jackson C, Best N, *et al*. *The BUGS Book: A Practical introduction to Bayesian analysis*, CRC Press (2012).

[2] Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual, version 2.27*. (2021) https://mc-stan.org — Accessed 9 August 2021.

[3] McElreath R. *Statistical Rethinking*, CRC Press (2016).

[4] University of Bristol. *Centre for Multilevel Modelling — Software*. https://www.bristol.ac.uk/cmm/software/ — Accessed 14 February 2022.

[5] Plummer M. *JAGS: Just Another Gibbs Sampler* https://mcmc-jags.sourceforge.io/ — Accessed 9 August 2021.

[6] Heitjan DF, Rubin DB. Ignorability and coarse data. *Annals of Statistics* (1991); 19(4): 2244–53.

[7] Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* (1999); 10(1): 37–48.

[8] O'Hagan A, Oakley J. SHELF: the Sheffield Elicitation Framework. http://www.tonyohagan.co.uk/shelf/

[9] Grant RL. *A philosophical standpoint for the practice of statistical inference.* http://www.robertgrantstats.co.uk/papers/response-sist.pdf – Accessed 14 February 2022.

[10] Nicenboim B, Schad D, Vasishth S. *An Introduction to Bayesian Data Analysis for Cognitive Science*, Table 16.1, quoting Jeffreys H, *Theory of Probability* (1939), Clarendon Press (*aka* OUP). https://vasishth.github.io/bayescogsci/book/hypothesis-testing-using-the-bayes-factor.html#tab:BFs

[11] Gabry J, Simpson D, Vehtari A, Betancourt M, Gelman A. Visualization in Bayesian workflow. *Journal of the Royal Statistical Society, Series A* (2019); 182(2): 389–402.

[12] Grant RL. *Data Visualization: charts, maps and interactive graphics*, Chapter 8, CRC Press (2018).

[13] Spiegelhalter DJ. Funnel plots for comparing institutional performance. *Statistics In Medicine* (2005); 24: 1185-202.

[14] Kunadian B, Dunning J, Roberts AP, *et al*. Cumulative funnel plots for the early detection of interoperator variation: retrospective database analysis of observed versus predicted results of percutaneous coronary intervention. *British Medical Journal* (2008); 336: 931–4.

[15] Gale J, Ooms A, Grant RL, Paget K, Marks-Maran D. Student nurse selection and predictability of academic success: The Multiple Mini Interview project. *Nurse Education Today* (2016); 40: 123–7. http://robertgrantstats.co.uk/papers/MMI.pdf

[16] Norris M, Hammond JA, Williams A, Grant R, Naylor S, Rozario C. Individual student characteristics and attainment in pre-registration physiotherapy: a retrospective multi-site cohort study. *Physiotherapy* (2018); 104(4): 446–52.

[17] Reed E, Todd J, Lawton S, Grant RL, Sadler C, Berg J, Lucas C, Watson M. A multi-professional educational intervention to improve and sustain respondents' confidence to deliver palliative care: A mixed-methods study. *Palliative Medicine* (2018); 32(2): 571–80.

[18] Grant RL. *Robert Grant - Stats - Publications* http://www.robertgrantstats.co.uk/publications.html#papers

[19] Song XY, Lee SY. *Basic and Advanced Bayesian Structural Equation Modeling*, Wiley (2012).

[20] Rubin DB. Inference and missing data. *Biometrika* (1976); 63(3): 581–92.