Bayesian generative modeling of student results in course networks

Haas, M.R.; Caprani, C.; Beurden, B.T. van

JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 10(3), 135–152. https://doi.org/10.18608/jla.2023.7957

Bayesian Generative Modelling of Student Results in Course Networks

Marcel R. Haas[1], Colin Caprani[2], Benji T. van Beurden[3]

Abstract

We present an innovative modelling technique that simultaneously constrains student performance, course difficulty, and the sensitivity with which a course can differentiate between students by means of grades. Grade lists are the only necessary ingredient. Networks of courses will be constructed where the edges are populations of students that took both connected course nodes. Using idealized experiments and two real-world data sets, we show that the model, even though simple in its set-up, can constrain the properties of courses very well, as long as some basic requirements in the data set are met: (1) significant overlap in student populations, and thus information exchange through the network; (2) non-zero variance in the grades for a given course; and (3) some correlation between grades for different courses. The model can then be used to evaluate a curriculum, a course, or even subsets of students for a very wide variety of applications, ranging from program accreditation to exam fraud detection. We publicly release the code with examples that fully recreate the results presented here.

Notes for Practice

- When modelling student results, the approach presented here can be used as a fully Bayesian statistical framework that is easy to adapt and extend.
- Strongly simplified models for learning outcomes constrain the difficulty of courses better than the performance of students and can therefore be effective and efficient tools for curriculum evaluation.

Keywords

Learning analytics, statistical modelling, bayesian generative modelling, curriculum evaluation

Submitted: 26/01/2023 — **Accepted:** 19/11/2023 — **Published:** 15/12/2023

[1] *Corresponding author Email: datascience@marcelhaas.com Address: Department of Public Health and Primary Care/Health Campus The Hague, Leiden University Medical Center (LUMC), Albinusdreef 2, 2333 ZA Leiden, Netherlands; Business Intelligence, University of Amsterdam, Spui 21, 1012 WX Amsterdam, Netherlands ORCID ID: https://orcid.org/0000-0003-2581-8370*

[2] *Email: Colin.Caprani@monash.edu Address: Department of Civil Engineering, Monash University, Melbourne, 23 College Walk, Clayton, Victoria 3800, Australia ORCID ID: https://orcid.org/0000-0001-6166-0895*

[3] *Email: b.t.vanbeurden@uva.nl Address: Methods & Statistics, Research Institute of Child Development and Education, Faculty of Social and Behavioural Sciences, University of Amsterdam, Postbus 15780, 1001 NG Amsterdam, Netherlands ORCID ID: https://orcid.org/0009-0004-9339-0710*

1. Introduction

The field of learning analytics (LA) is broad and often captures detailed cognitive processes by measuring engagement and the results of small incremental steps in the learning process. This can happen in online learning environments, where detailed behaviour can easily be followed at high temporal resolution. To evaluate the learning progress and drivers thereof within one "course" (i.e., a limited set of learning goals), this works well, but at the scale of entire university curricula this becomes very cumbersome and computationally difficult to assess.

Predictive analytics, where a particular target variable is predicted based on predictor variables (or features), is often used on an individual student basis (Strecht et al., 2015; Jeffreys, 2015; Asif et al., 2017; Daud et al., 2017; Cui et al., 2019). With predictors coming from both the previous study results and engagement, as well as demographic data, the results of a particular student for a particular course or exam can be modelled as a fairly straightforward supervised machine learning (ML) problem. Such forecasts can be used for intervention during a course (Rienties et al., 2017), for offering additional educational help in the near future (F. Chen & Cui, 2020), or for predicting dropout (Archambault et al., 2009; Lacave et al., 2018); many more applications can probably be thought of.

Even though interesting in itself, individual results are short-term one-off predictions, but when combined they offer the possibility to zoom out from the individual learner to a class of students (Schmitz et al., 2022), an educational program, or even an entire educational institution. LA, even when the model is student based, can readily be used at the course or curriculum level. In fact, due to the scale-up, uncertainties are likely to decrease. Indeed, there have been many efforts to model learning progression, including knowledge retention and forgetting (Sense et al., 2021), as well as to model the flow of students through a curriculum, e.g., using agent-based modelling (McEneaney & Morsink, 2022) or by modelling the cohort as the sum of its parts (Shah & Burke, 1999; Nicholls, 2007; Raji et al., 2021; Munguia & Brennan, 2020). Other efforts take a helicopter view of the curriculum and propose to model it as a whole (Ochoa, 2016). Modelling alone is not usually enough to in fact improve the educational program (Macfadyen & Dawson, 2012), but it leaves room for evidence-based LA (Ferguson & Clow, 2017) and effective stakeholder management (Dawson et al., 2018).

The statistics of much of the existing body of quantitative modelling work has been frequentist in nature. Like many quantitative fields, LA has moved from dashboarding with descriptive statistics to often ML-based predictive modelling (a.k.a. PLA) at a high pace in recent years (Cui et al., 2019). Uptake of more modern methods in LA has previously been slow, likely in part due to lack of understanding of the methodology and its use cases (Dawson et al., 2018), but also because of the varied ways in which course and curriculum design are implemented at various institutions; the different involvement of the various stakeholders (Ferguson & Clow, 2017; Rienties et al., 2017); and the lack of a common evaluation of LA methods, as outlined by Herodotou and colleagues (2019).

Sitting between descriptive statistics and black box–like modelling like many ML applications, there is the realm of latent variable modelling, in which knowledge of the real world can easily be incorporated, and after which the model parameters in fact mean something about the phenomenon modelled. Especially for LA, where the hidden or latent variables often describe properties of the educational material or the students, discovering latent parameters that connect straightforwardly to the domain of application is very powerful. As Gardner and Brooks (2018) noted, model methodology and evaluation are better done in the Bayesian realm than with frequentist statistics, and using a model that is Bayesian in nature is thus a straightforward choice. Not many LA studies have used Bayesian methods so far. Notably, Lacave and colleagues (2018) use a Bayesian network to predict student dropout using many academic as well as demographic features, partly based on the work of Di Pietro and colleagues (2015), who use a Bayesian network of ordinal and nominal variables to monitor a master's program and point out critical issues. The former unfortunately remark that detailed statistical knowledge is not necessary when software packages are used for the modelling, bringing the application into the same black box–like domain as the ML-based models alluded to before.

In this paper, we propose a new LA framework that models learning progression holistically (i.e., from the scale of single student grades per course up to program curricula or entire universities). As its input, it requires a data set with a grade list per course. For each course, the framework uses Bayesian generative modelling to simultaneously estimate a difficulty, based on the grade distribution and the performance of the student population, and a sensitivity, based on the capacity of a course to distinguish between low and high performance. We model and parameterize the full data-generating process and estimate the parameters of the model by sampling their posterior probability density functions (PDFs) using standard methods from Bayesian statistics. As such, this is a first and very simple fully Bayesian *generative* LA model that can contribute to meeting the existing challenges described above, from predictive modelling of student progress to the evaluation of curricula and the educational offerings of an institution. The generative nature implies that latent variables have a clear meaning, aiding adoption by stakeholders and enabling actionable insights for teachers, educational researchers, and policy-makers. It is a fully open model, and therefore its inner workings and parameters are clear to any user, which aids future extension of the model. Despite its simplicity, the model is used in practical applications as well, and the results prove to be useful and promising.

In Section 2.1 we will outline how student performance, course difficulty, and grades depend on each other in our model. Section 2.2 outlines the build-up of networks and Section 2.3 describes the modelling technique used to constrain latent parameters based on observed grades. The ethical component and potential misuse of our approach are discussed in Section 2.4. We will also show how this all works in practice, using some idealized experiments based on simulated data, in Section 2.5. In Section 3 we show the resulting models and statistics about courses and students for a real and open data set from the Open University, illustrating some of the requirements that must be met by the data set in order to model these successfully. In Section 4 we will model a data set from the University of Amsterdam College of Law to show that under realistic conditions, these techniques give actionable results. A discussion of the practical use of these models follows in Section 5, after which we conclude in Section 6. With this paper we release a public code repository[1] with example data. In there, all figures from this paper are reproduced and it is illustrated how the code can easily be used on one's own data and how the model can be extended for more sophisticated and realistic use cases.

[1] https://github.com/harcel/BGM_CourseNetworks.

2. Methods

In this section, we describe our modelling approach, which consists of three main ingredients. First of all, we employ a very simple relation between properties of students, of courses, and of the grades obtained. We can only consistently model these if the courses form a network through overlapping student populations attending the courses, which is the second ingredient. Finally, if they do, we can construct a Bayesian generative model in which we constrain course difficulty, course sensitivity, and student performance simultaneously from the list of grades.

2.1 A Simple Model for Grades

We build on a model for study results of students that starts from a very simple assumption that we borrow from Item Response Theory (IRT, e.g., Embretson & Reise, 2000). In IRT, which is commonly applied to right/wrong or Likert-scale questions within one test, the probability of correctly answering a question on a test depends on the student's ability (α) and the question's difficulty (δ). In fact, the log-odds of answering correctly is commonly assumed to depend linearly on the difference between the two ($\alpha - \delta$). The probability is, then, given by the logistic function, i.e., $P(\text{correct}) = 1/(1+\exp(-(\alpha-\delta)))$. A term for the sensitivity (s, also called *discrimination*) of a question to distinguish between correct and incorrect responses (the sharpness of the transition of $P(\text{correct}) = 0 \to 1$) can be added in the exponent.

We adopt and modify this strategy here and model grades for courses as a function of student performance, course difficulty, and course sensitivity. We use the terms *performance* (analogous to *ability* in IRT) and *difficulty* very loosely here. The difficulty of a course reflects the probability of getting a high grade. Sometimes, very difficult courses may end in a relatively easy exam, pushing what we call *difficulty* strongly down. Conversely, a very easy course may be assessed by a strictly corrected essay, which could make the course seem more difficult in the results. Also, performance of students is not a simple one-dimensional concept, of course. Here, though, we use the term broadly to describe the probability for a student to score a high mark for the given testing method for the courses attempted.

We adapt the IRT approach to be able to capture continuous grades for tests or courses that depend similarly on student performance, course difficulty, and perhaps a course sensitivity. Not all grading systems are on the same scale. Here we will assume that the scale is numeric and runs from a minimum grade $G = g_{\min}$ to a maximum grade $G = g_{\max}$, following the relation

$$g = \frac{g_{\max} - g_{\min}}{1+\exp(s\cdot(\delta-\alpha))} + g_{\min}, \tag{1}$$

in which the numerator and the extra term set the scale of the grades. We denote this by a lowercase g to separate it from the actual grade obtained, G, i.e., the above "ideal" grade g with some scatter, since students may have a good or a bad day, get lucky or unlucky, etc.: $G = g + \varepsilon$, in which ε is scatter that can have any functional form, as long as $G \in [g_{\min}, g_{\max}]$.

The grades here should be as close to "raw grades" as possible: sometimes grades are adapted either to not be close to pass/fail boundaries or to increase a class average or passing percentage to an acceptable level. As discussed below, some such policies for handling grades can be easily incorporated in the modelling process, while others can't.

Note that performance and difficulty are on an arbitrary scale in the same arbitrary units. It is only the difference that matters: $\alpha - \delta = 0$ corresponds to a grade halfway along the scale. For a graphical representation of dependence of grades on the constituent variables, see Figure 1, where we use a grade scale that runs from 0 (worst) to 100 (best), in which a grade of $G > 40$ is typically considered a pass, like in the data set presented in Section 3. Grades can be continuous or discrete but typically have integer resolution.

2.2 Networks of Courses

Networks have recently become more popular in LA (B. Chen & Poquet, 2022). Networks come in many varieties, such as social, communicative, or collaborative, with strong potential influence on the learning process. Network analysis provides LA practitioners with descriptive, analytic, and evaluative tools, through quantitative measures of network properties, as well as the visual aids that networks provide.

The validity of network analysis results in LA critically depends on network definition choices (e.g., Wise et al., 2017). Here, we keep our network definition very simple; it only serves as a tool to see whether the data set at hand has an adequate structure for further modelling, and when necessary as a tool to select the largest subset of data that we can use. We construct the network as follows.

We build a network of courses in which the nodes are courses and the edges are overlapping student groups: two nodes are connected by an edge if at least a minimum number of students (a free parameter, in our experiments set to 5) finished both courses[2]. When a student finishes a course, there will be a grade for that course, and, as such, it is trivial to construct this

[2]Technically, we construct the network as follows. We build a network by first connecting every student with every course that the student took. As such,

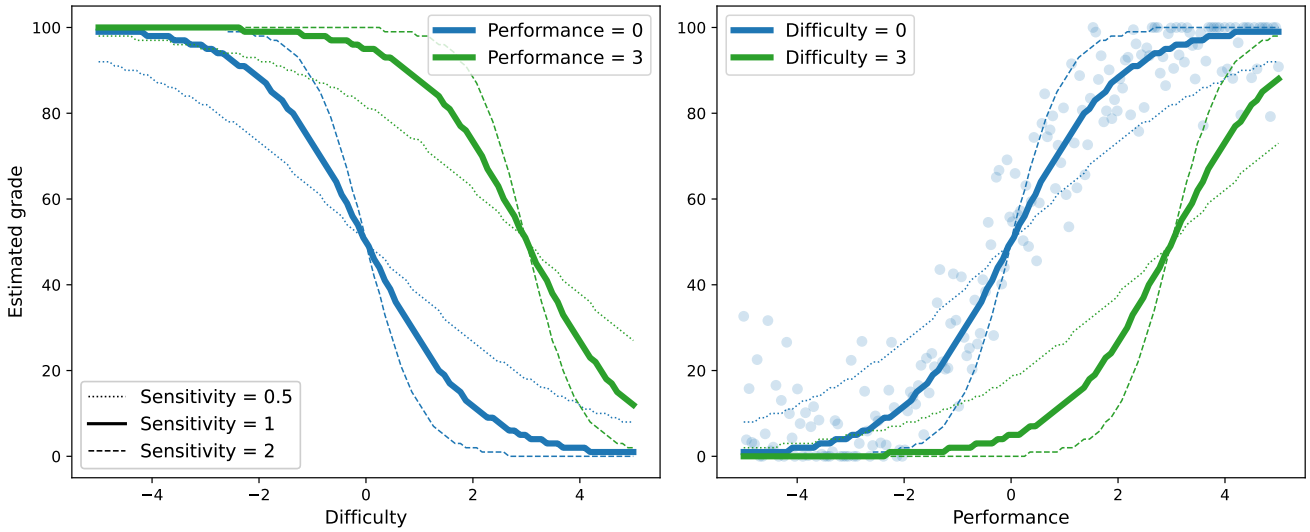

Figure 1. The dependency of grades, on a scale from 0 to 100 on difficulty and performance. In the left panel, student performance is fixed at 0 (average) and 3 (very high), so the grade is a declining function of difficulty, with steepness increasing with sensitivity. In the right panel, difficulty is fixed, so the grade is an increasing function of student performance, with steepness increasing with sensitivity. The scattered points in the right panel are randomly sampled grades with considerable scatter (a standard deviation of 12.5 on this scale), as used in the experiment described in Section 2.5. The clipping of grades to the range [0,100] is clearly visible by the piling up of grades at the upper limit.

network of courses based on grade lists alone. These networks serve as an auxiliary tool, from which it is easy to select the largest subset of courses for which we can analyze the results.

Students take courses. Their performance is reflected in their grades, but only once the difficulties and sensitivity of these courses are known. The reverse is also true: courses are taken by students, and the difficulty of a course is reflected in the grades of the students. We can only constrain the difficulty of courses and the performance of students in a fully connected network. If there is no link between two courses or two populations of students, their combinations of difficulty and performance cannot be linked. There needs to be an exchange of information between all elements in the network. We will from now on focus on only one interconnected network of courses. As an illustration, the left network in Figure 2 cannot be analyzed as a whole but only for the upper and lower sub-networks, respectively. The right panel in that figure shows how one extra course taken by (a subset of) the students from the upper and lower networks results in one connected network that can be modelled as a whole. If that "Course X" were not present, the two groups could be modelled separately, but one group could consist of students with a higher (or lower) performance, following more (or less) difficult courses, respectively, than the other group and it would not be possible to distinguish between them.

Note that this approach neglects any time dependence of the curriculum progress. Time dependence in the models we will describe below can be incorporated but is outside the scope of the current work. For data with many data points close together in time, such as following along with MOOC education or the logs of a digital learning environment, this aspect is crucial (F. Chen & Cui, 2020). Here we need only assume that students have mastered the prerequisites, which should be handled by the registration process. We will come back to incorporating time dependence in Section 5.

2.3 Bayesian Generative Modelling

2.3.1 A Model for Student Results

A generative model is a model in which we describe the data-generating process in a probabilistic programming language (McElreath, 2020). The data-generating process here is as simple as described above: students have a performance, courses have a difficulty and a sensitivity, and we assume that the dependence of the observed data (a list of all grades obtained by all

the preliminary network consists of nodes that are students and courses, an edge between which indicates that the student at one end took the course at the other end of the edge. All neighbours of all student nodes are courses (the set of courses the student took), and all neighbours of all course nodes are students (the students who took the course). Therefore, all second-degree neighbours of a course are courses that were taken by students that took both. This allows us to finally construct a course network by connecting all course nodes to their second-degree neighbours to end up with a network in which all nodes are courses and the edges between them denote the overlapping student population (the number of length-2 paths from course A to course B is the size of the population of students that took both courses).

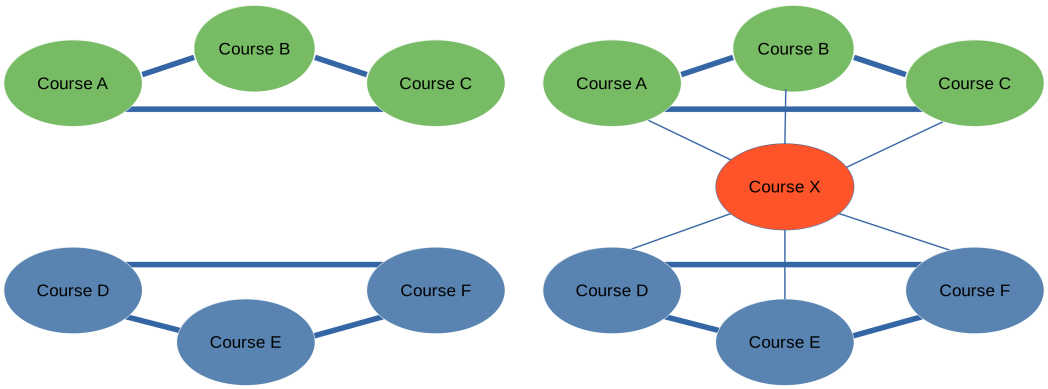

Figure 2. An example of small networks of courses. *Left:* A network that consists of two separated networks (denoted by the two colours), in which it is impossible to simultaneously constrain performances and difficulties, as explained in the main text. *Right:* The same networks, but now connected through a course (orange) that links the two. The thin lines need not all be there, as long as there is at least one connecting to the lower network and one to the upper network.

students for the courses they followed) is described by Equation 1.

Conditioning the model on the observed data constrains our knowledge about the model and its constituents. The method is Bayesian and hence results in a *posterior PDF*, or *posterior*. This is a PDF for all model constituents under the assumption that the model itself is valid, having observed the data. It is denoted by $P(\text{parameters}\,|\,\text{data})$, which can be read as "the probability of the model parameters, given the observed data." This posterior depends on a likelihood (which is a measure of the probability of observing the data for a given model parameter choice) as well as previous knowledge about the model parameters (which can be constrained by theory or by previous work, or can be very unconstrained) by Bayes' equation:

$$P(\text{parameters}\,|\,\text{data}) \propto P(\text{data}\,|\,\text{parameters}) \cdot P(\text{parameters}), \tag{2}$$

where the constant of proportionality $(1/P(\text{data}))$ is called the marginal likelihood and is fixed for a given data set and likelihood function. The likelihood function we choose here is

$$G \sim [\mathrm{N}(\text{data}\,|\,\mu = g(\alpha,\delta,s), \sigma = \varepsilon)|_{g_{\min}}^{g_{\max}}, \tag{3}$$

which reads like "the observed grade is a normally distributed stochastic variable with a mean (μ) given by $g(\alpha,\delta,s)$ (i.e., Equation 1) and a unknown standard deviation (σ) of ε, truncated to range between minimum and maximum grades." When grades close to pass/fail boundaries are rounded more coarsely than on other parts of the grade continuum, this can be incorporated here as well, making the functional form for G somewhat more complicated, but the modelling described below is still identical. Also, it is not necessary to opt for the normal distribution; in fact any distribution can be used instead with only very minor modifications to the code.

2.3.2 Choice of Priors

In any Bayesian analysis, the choice of prior distribution functions for model parameters is crucial. They reflect prior knowledge of the model parameters that subsequently gets updated with the observed data to form the posterior knowledge of the model parameters. Priors can come from previous work, theory, or intuition and can be very well constrained in some cases (very narrow PDFs) or very uninformative, meaning that a large range of values for the model parameters are about equally likely.

The right-most term of Equation 2 is the collection of prior PDFs, which in the language of probabilistic programming can be written as

$$\begin{aligned} \alpha &\sim \mathrm{N}(0,3), \\ \delta &\sim \mathrm{N}(0,3), \\ s &\sim \mathrm{LogN}(0,0.25), \\ \varepsilon &\sim \mathrm{N}^{+}(0,1), \end{aligned}$$

which, together with Equation 3, defines the model. All the right-hand sides are known PDFs: $\mathrm{N}(\mu,\sigma)$ is the normal distribution with mean μ and stand deviation σ, LogN is the log-normal, and N^{+} is the positive half-normal (in this case, only the positive

half of N(0,1)). The priors in this work are all chosen to be uninformative (i.e., a wide range of values is allowed), but we did constrain the standard deviation of the likelihood to be positive and the sensitivity to be positive and centred around one. A graphical representation of the computational graph, including the dimensionality of the experiments described below, is given in Figure 3.

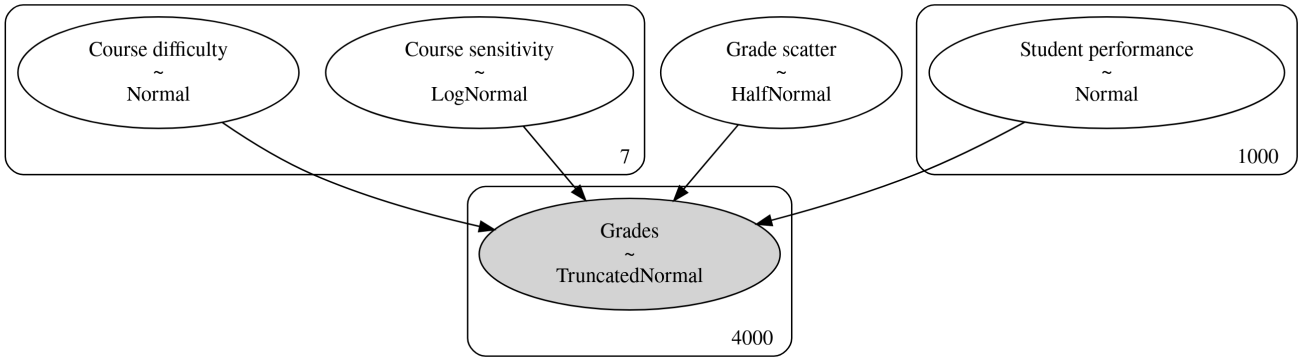

Figure 3. The computational graph for modelling the observed grades is non-hierarchical. The numbers indicate the number of computational elements: there are seven courses, 1000 students, and 4000 grades in the main example discussed in Section 2.5.

2.3.3 Monte Carlo Markov Chain

The posterior in general is hard to compute and will therefore be sampled with a Monte Carlo Markov Chain (MCMC), which results in a number density of sampled points on the chain that is directly proportional to the local probability density (e.g. Diaconis, 2009). We employ the No U-Turn Sampler (NUTS, Hoffman & Gelman, 2014), which is a Hamiltonian sampling algorithm (Neal, 1994). The result is a so-called trace with all samples (one sample on the trace has a value for each of the parameters in the model), of which summary statistics can be used to summarize the model results. In the idealized experiments described in the next section, we will plot the marginal distribution functions of parameters of interest for every separate Markov Chain, as well as a representation of the trace itself, for diagnostic purposes. In the remainder of the paper we will combine the posterior samples of separate Markov Chains into one posterior PDF and we will omit the visualization of the trace. The online code repository does reproduce them for the interested reader.

2.4 Ethics and Potential Misuse

We fully realize that our one-dimensional representation of student performance doesn't do justice to the complex phenomenon of cognitive ability or aptitude to particular skills. At the scale of an entire university or curriculum, grade fluctuations are potentially a consequence not only of student-specific characteristics like intelligence, domain-specific knowledge, and motivation, but also of other factors, like the quality of the provided education and the type of assessment. While extending the methodology from IRT, we deliberately changed the wording from *ability* to *performance* to indicate that what we attempt to constrain here is a narrow view of what one would call ability. We strongly discourage the use of the performances derived with the models below as a measure of student intelligence, a predictor of student success (either within the program modelled or outside their academic studies), or a property of people that can be independently measured. It merely serves here as a necessary latent variable that allows us to say something about courses in a curriculum (or a collection of interconnected curricula), and that the model is likely to measure something useful about the courses, given the grades and latent variables when performances are well constrained (i.e., on a per person basis, they are well determined with an uncertainty per student much smaller than the scatter between students). We will not report or draw conclusions on student performance for these reasons.

In the triangle of assessment, epistemology, and pedagogy (Knight et al., 2013), models like these tend to draw the user too much to the corner of assessment, away from epistemology and pedagogy, even though the latter two are arguably the phenomena that are to be measured (Lang et al., 2017). Therefore, it is important to keep in mind that even though these are quantitative and flexible models, they only capture a fraction of the cognitive process students are going through in an educational program.

2.5 Idealized Experiments

In order to test our model, we set up experiments using simulated data. We create the network that is shown in the right-hand panel of Figure 2. Five hundred students, with a variety of "performances," attend courses A, B, and C. Another 500 attend D, E, and F, who also have a variety of performances but significantly shifted from those of the former group. The courses are

graded between $g_{\min} = 6$ and $g_{\max} = 10$. The courses are also different in difficulty so that the distributions of grades for the two groups of students are the same, with average grades for the three courses being 7.4/7.5, 7.7, and 8 for both groups of students. All attend Course X, the linking course between the two networks. Because that course has a fixed difficulty, the grades from the students of the two groups for Course X differ substantially: averages of 7.0 and 8.6, respectively, resulting in an average grade for Course X of 7.8. The standard deviations of grades of all courses, including Course X, are in the range 1.19–1.38.

Looking at the average grades for all courses as well as their standard deviations would not at all suggest that there is a difference in the two sub-groups of students, and without noticing the bimodal distribution of grades in Course X (which in a curriculum review is easily overlooked), it would be hard to identify.

The grades are calculated with a sensitivity parameter equal to 1. The scatter around the expected grade g (ε) is drawn from a normal distribution with 0 average and a standard deviation of 0.5, which is large compared to the range of grades (6–10) obtainable. Note that when grades go below 6 or above 10, they are clipped at 6 and 10, in both data generation and the generative model. An analogue of this amount of scatter around the expected grades was shown in the right panel of Figure 1, where the scatter is also 12.5% of the total range of grades.

When these grades are modelled with the models described above, the posterior traces and resulting posterior distribution functions for the quantities related to the courses are shown in Figure 4. There are too many posterior distribution functions for the students to show, so we will omit those. Their performances are well recovered with very small uncertainty. The trace plots (right-hand panels) look somewhat like caterpillars, which is what we want. The bands are horizontal, so the region of parameter space walked by the chains is not moving, which indicates that the model has converged (the converse behaviour indicates non-convergence). The trace mostly explores high posterior density regions and sometimes moves a bit further out, toward the outskirts of the high probability density regions. The PDFs in the left-hand plot are all shown by four different line styles in the same colour. These are four independent chains, which should largely agree on the posterior PDF.

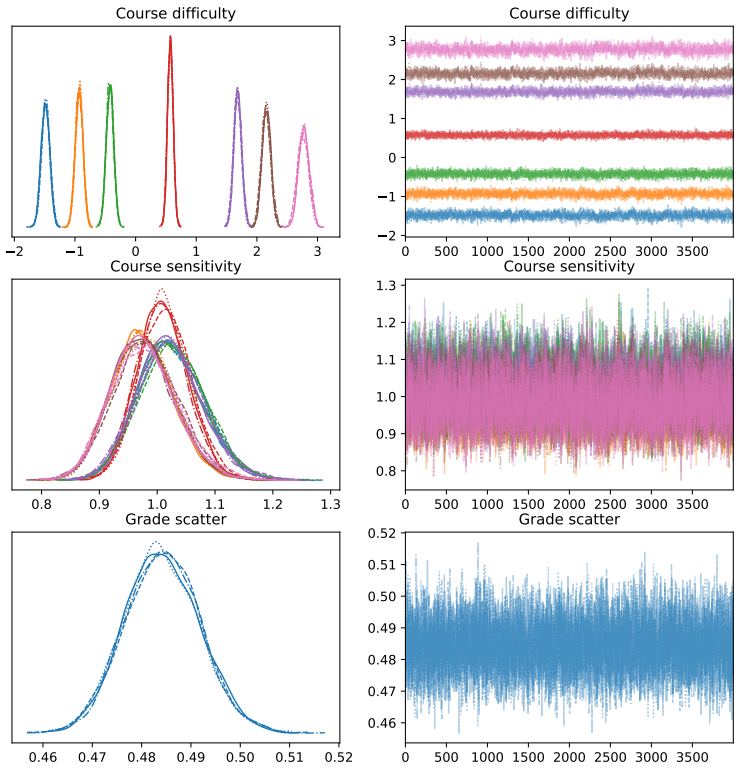

Figure 4. Posterior PDFs (*left*) and traces (*right*) of model results on simulated data.

In these models, the average performance of students is forced to be 0. The average difficulty is then set by the scale of the

grades, because an average grade halfway along the scale (in this case, 8) would correspond to an average difficulty of 0.

The recovered sensitivity is very close to the input value of 1. Here, the necessity of using the truncated normal distribution in the generative mode shows: when modelling the likelihood with an untruncated normal distribution, the recovered sensitivity would be close to 0.7. This is a result of the clipping of the grades. When a grade goes below 6 or above 10, it is limited to that range. Therefore, at the low end of the grade range, grades will only bias upward, and the reverse is true at the high end of grades. This pushes the average grades toward a shallower function of increasing performance (or decreasing difficulty); c.f. Figure 1. It is confirmed that, indeed, when repeating the experiment with much less scatter around the expected grade, the sensitivity moves upward toward the input value of 1, even when modelled with an untruncated normal. Because we take the truncation into account while conditioning on observed data, we do not have to worry about artificial flattening of the dependence of grades on performance and difficulty.

We have repeated the experiment with different settings and conclude that, as long as the grades are indeed described by the relation with which we model them, we have the following results:

- Course difficulties are very well recovered, even if the variety in performance becomes very large.
- The number of students following Course X does not need to be big; even with just 50 students (10%) in that course, the recovery of difficulties and performance is still very good, with only slightly larger uncertainties. For even lower fractions of students in that course, uncertainties on all recovered performances and difficulties go up. This shows that only small information exchange between well-connected sub-networks is enough for inference.
- The scatter on the grades can be made larger, resulting in only marginally higher uncertainties on recovered model parameters.

2.6 Computational Demand

Bayesian analysis with MCMC is known to be computationally intensive. Networks of courses can get big for large data sets, and using uninformative priors can lead to long sampling times of the Markov Chains. The models we present here are still relatively cheap and are all fully sampled on a moderate laptop with a 12-core i7 processor. The idealized experiments described above sample in just a few minutes, while the more realistic data set described below in Section 4 takes up to a few hours on that same hardware. Sampling time scales with the number of parameters that need to be estimated, so with a large number of courses or students (the numbers used are quoted here), this can grow. In general, it would be wise to include only the model parameters that are needed to constrain the problem of interest, and with the model set-up used in this paper, personal hardware should be enough even for large institutions, modelling up to a few years' worth of data.

3. Modelling the Open University Public Data Set

Data that are not generated under the assumptions of the model, but rather in the real world, are crucial to illustrate the applicability of this modelling scheme. The Open University in the United Kingdom released a data set for LA (Kuzilek et al., 2017). In this data set, a large number of anonymous student results are recorded for both assignments and exams. The assignments come with weights that add up to 100, but it is not clear how the total result of a course depends on the assignments and the exam. Also, many students have in fact only done some assignments and certainly not always the exam.

3.1 A Loosely Connected Network

Because we would end up with a very small network of interconnected courses otherwise, we decided to split every course into two parts, each with its own difficulty and its own population of students: the exam is one, and the collection of assignments is the other. *These two parts are considered separate courses in the rest of this section.* Indeed, these forms of assessment typically have quite different outcome characteristics (Richardson, 2015). For the assignments, we take the weighted average of the grades, whereas an exam has just one result: the grade. Note that for the assignments, we do not set the grade for an assignment to 0 if there is no recorded grade for a given student for a given assignment, but rather we take only the actual recorded results into account. Therefore, the sum of weights is not always 100, but it is the sum of the weights of recorded results for that student. Courses in the data set are called AAA, BBB, etc., and we use the same designation here for the exam results (for course BBB there are no exam results in the data set), while the course name is appended with "_a" for the assignments. We do not take into account any other piece of information from the data about students, courses, or assignments.

The course network is shown in Figure 5 and consists of seven "courses." The number of students ranges from 1915 to 6033 and averages at 3863 per course, while the total number of students is 20,424. The grades run from 0 to 100, so when modelling we relax the prior on ε to an uninformative $\mathrm{N}^{+}(0, 10)$. The distributions of grades, as shown on the diagonal of Figure 6 below, are nowhere near similar in shape: some are strongly peaked at different peak values, while others show very flat distributions, indicating perhaps a large spread in sensitivity, or very different population distributions of student performance. We relax the prior on the sensitivity parameter to $s \sim \mathrm{LogN}(0, 0.5)$.

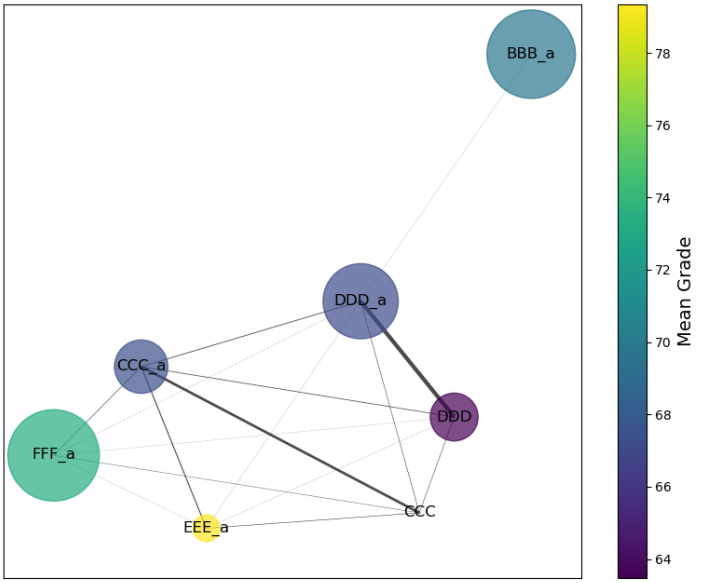

Figure 5. The course network of the Open University data set. Exams are denoted by just three letters; the assignment sets are appended with "_a". For course BBB there are no exam results in the data set. The thickness of the edges in the network (only shown if at least five students have a result in both nodes) indicates the overlap: strong between the assignments and the exam of the same course, but much weaker between courses (typically a few percent of students overlap between courses). The sizes of the nodes in the network denote the numbers of students, which range from 1915 to 6033, while the colour indicates the average grade.

3.2 Signs of Unsatisfactory Data Properties

Comparing the numbers of students per course and the total number of students in the data set already indicates that there is in fact very little overlap in student populations of courses in this data set. Indeed, the number of students that appear in more than one course is large for those courses for which both assignments and an exam are in the data (3037 and 1915 for courses DDD and CCC, respectively); for the other courses the overlapping population ranges from six (the imposed minimum was five) to 722, with an average of 240 (1.2% of the population). Note also that the course with only one connection to the rest of the network (BBB_a) is connected with 10 students out of the 5769 that took the course. This is expected to result in large uncertainty in the recovered parameters for this course. The posterior of networks like this takes too long to converge and sample. With the limited information exchange between BBB_a and the other, more strongly interconnected, nodes of the network, we decided to disregard BBB_a for further analysis.

In Figure 6 we show a detailed exploration of the distributions of grades and the correlations between grades in different courses for the overlapping student populations. When modelling with the techniques described in this paper, there should be considerable correlations between grades (but see Section 5 for directions for future work in which this no longer needs to be the case). Various correlations in this data set are weak or absent. Also, some courses have very small overlapping student populations, which limits information exchange between the nodes in the network and thus inhibits precise determinations of course difficulties and hence student performance.

Indeed, when modelling this data set with the same set-up as in Section 2.5, the model has difficulty converging (i.e., it does not easily find a stable posterior PDF). It is possible to tune the sampler and find a solution that according to all numerical criteria would converge, but that this is necessary is usually a sign of inferior model specification. Especially in the case of this data set, the model might not be fully appropriate. Besides the very limited information exchange due to the thin connections in the data set (which might call for a hierarchical model structure that ensures pooling over the courses, so they do not diverge too much), we also see a hint of a challenge in the grade distributions as depicted along the diagonal of Figure 6. If courses are assigned one given difficulty and students have a fixed distribution of performances, then the grade distributions are expected to be shifted and skewed/compressed versions of one another. In this case, we see strongly peaked as well as very broad

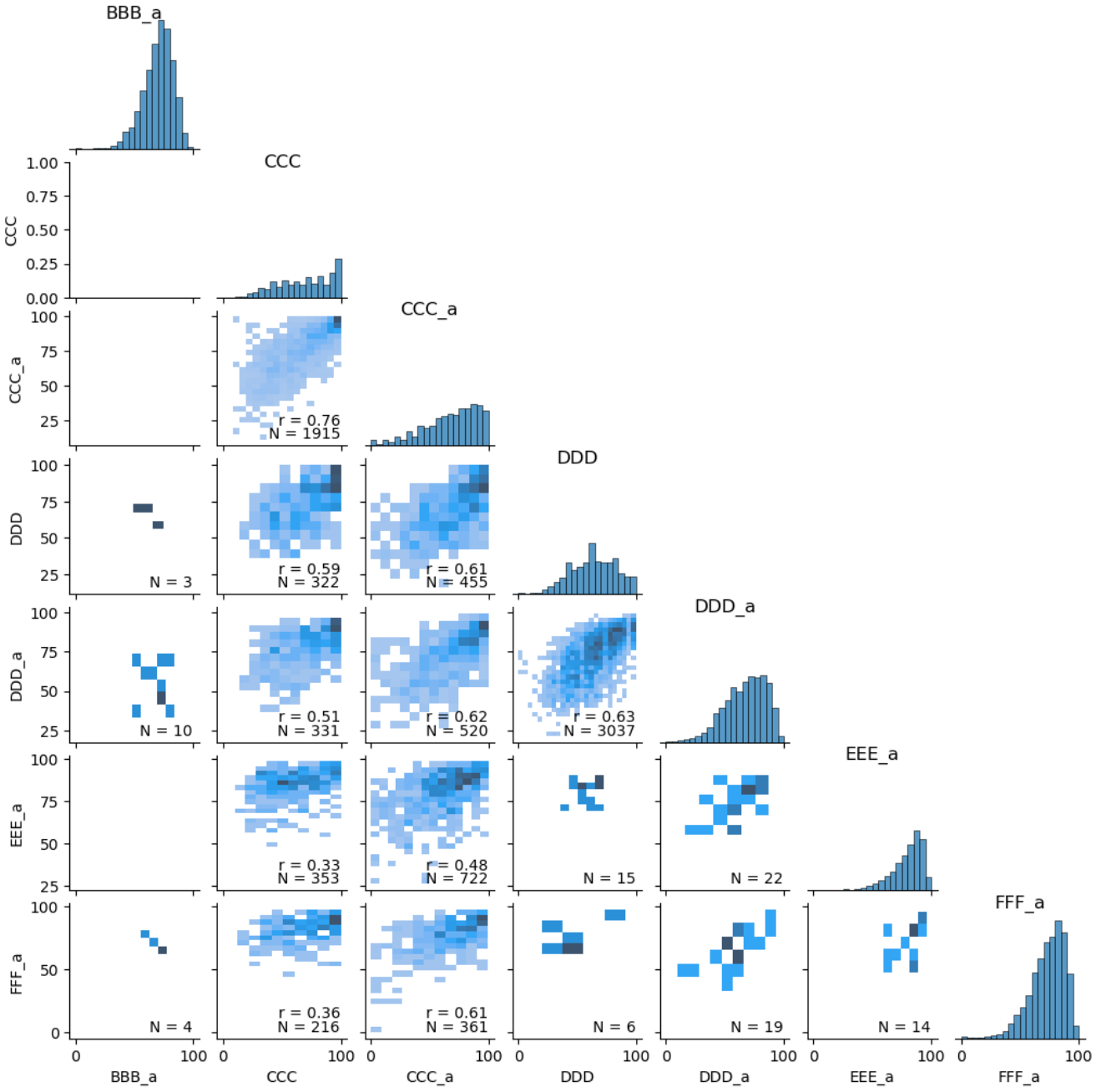

Figure 6. An exploration of the distributions and correlations of grades for the different courses in the data set of the Open University. Along the diagonal, we show a histogram per course, with bins of width five. Under the diagonal we show a two-dimensional histogram, with the number of students that have a grade for both "courses" indicated and, if that number is higher than 30, the correlation coefficient between the grades.

distributions. This hints at potentially very different sensitivities, but if this is combined with a very small overlapping student population, the posteriors become very hard to constrain: is the student population in one course very different from that in the other, is the sensitivity of one course very different from that of another, or both? With a very low sensitivity (which translates loosely into the same grade for everybody, regardless of student performance), the difficulty can be very far from other difficulties. Depending on the average grade ($\alpha - \delta$), this results in everybody getting a score of 50, and in order to get to a typical score of 80 with very low sensitivity, the difficulty of the course needs to be very low, since the slope between grade and ($\alpha - \delta$) is shallow. Besides, course CCC has a large group of students piling up at the very maximum obtainable grade, which will make it difficult to use those grades to determine difficulty or performance.

Table 1. The numbers of students of the four best connected courses in the network and their overlap. All of the students who did the CCC exam also have a grade for the corresponding assignment set; for course DDD, seven students have an exam result but no assignments. Overlap is at least 322 students, or at least 7%.

Course 1	Number of students	*Course 2*	Number of students	**Number** and percentages of overlapping students	
CCC	1915	CCC_a	3318	**1915**	(100, 58)
CCC	1915	DDD	3044	**322**	(17, 11)
CCC	1915	DDD_a	4696	**331**	(17, 7)
DDD	3044	DDD_a	4696	**3037**	(100, 65)
DDD	3044	CCC_a	3318	**455**	(15, 14)
CCC_a	3318	DDD_a	4696	**520**	(16, 11)

The scatter on the grades is the one parameter that is hardest to constrain, as long as it modelled as a single parameter. Giving each course its own scatter, where all the scatters derive from a parent distribution (in a hierarchical model), helps, but not much. The online code repository gives an example of how to modify the model to do this.

With the considerable scatter, the very small overlapping student populations, and the highly dissimilar grade distributions, it is hard to constrain much about the data set at hand. The difficulty in converging is a red flag for interpretation of the inference.

3.3 A Well-Behaved Sub-network

In order to constrain course properties of some of the better-connected courses in the Open University LA data set, we select the four courses in our network that are most closely connected: CCC, CCC_a, DDD, and DDD_a. The numbers of students per course as well as per connection are given in Table 1.

This selection of courses still requires a model that uses a distinct scatter of the grades for each course. It results in some courses with very low sensitivity (0.2), which means that the course hardly distinguishes performance by grades and most of the variation in the grade must come from scatter caused by factors that are not modelled.

When also disregarding CCC (with its undesirable pile-up of students with a grade of 100), the model is much easier to tune, but with only three out of the original eight courses left, one could question the applicability of these models for the Open University data set. We conclude that these models require a data set with significant overlap in student populations between courses, grade distributions that do not pile up on either end of the scale (i.e., are calibrated in difficulty to actually differentiate between student performances), and at least some correlation between grades for different courses. As indicated, see Section 5 for suggestions on how to relax that last requirement.

4. Modelling Student Results at the Amsterdam College of Law

The Amsterdam College of Law teaches a range of bachelor programs at the University of Amsterdam. There are obligatory courses for all students, as well as electives that can belong to the same program or not. Students therefore can follow a wide range of curricula, in which a small part will be taken by all or most. We constructed a data set of all students at the College of Law in a bachelor program in one academic year. Due to re-takes or electives taken from another year in the program there can be overlap between the junior, sophomore, and senior cohorts. We only take students into account who pass the course. This is far from ideal, and we would like to stress the importance of pure and complete registration of grades. In this data set, for students who failed courses, the registration seems very incomplete (with exam failures and no-shows being indistinguishable in some cases, and no information about grades in other cases). When we construct the network as described in Section 2.2, we find a collection of multiple separate networks, which are not connected by an edge consisting of at least 10 students. Of these, we select the largest fully connected sub-network, which consists of 30 courses taken by between 108 and 858 students and with average grades ranging from 6.5 to 7.4 out of 10. Note that the minimum score in this data set is 6 out of 10, since that is the minimum for a pass. Figure 7 shows a graphical representation of the course network.

4.1 Modelling a Well-Connected, Large Network

The network shows a highly connected structure of about half the courses, where the course connecting student population is very big. Toward the outskirts of the network, the courses are attended by fewer people, so the overlap of students is also lower. Some courses are connected to almost every other course, while other courses have considerably fewer connections, because these are parts of smaller programs within the same faculty. The well-connected nature of this network makes it well suited for

Figure 7. The largest fully connected sub-network of all courses at the University of Amsterdam's College of Law. The nodes are the 30 courses, with the size corresponding to the number of students who followed the course and the colour indicating the average grade (dark is low; yellow is high). The thickness of the edges indicates the size of the overlapping student population (i.e., how many students attended both courses).

our approach, even though there is a big variation in the background and career paths of the students making up this network. This is essentially the opposite of the very sparsely connected network of the Open University in the previous section.

When modelling this network we need results for all courses. One course did not have a grade but instead only registered a pass or fail. When taking this course into account with a single grade for everybody (e.g., 6 for everyone, or the average of the whole data set for everyone), we find that the model cannot constrain the difficulty for this course at all. Firstly, when students of all performances obtain the same grade, the difficulty gets a very high uncertainty, and secondly, the grade assigned to everybody who passed the course sets the average difficulty retrieved by the model. Although this renders the determination of the difficulty of this course useless, it does strengthen our faith in the modelling technique, since this is precisely the behaviour expected from the model and its ingredients. It is clear, however, that for a careful evaluation of a course and its contribution to a curriculum, it is important that assessments of students happen on a common scale, at least if models like these are to be employed.

After the removal of this one course, the retrieved difficulties of all other courses remain unchanged, so there is little sensitivity to this source of noise in the data set. There are now 1985 students and 29 courses.

We model this population with the same model as we used for the idealized experiments in Section 2.5, depicted in Figure 3. It converges well without having to differentiate the scatter of the grades around the expected grade over courses, like we needed to in the Open University data set in Section 3. The scatter is somewhat large in comparison to the range of grades: the posterior mean of the scatter (ε in Equation 3) is at 0.83 (range of grades is 6–10). Still, all difficulties, sensitivities, and performances seem well determined, as displayed in Figure 8, where we compare the uncertainty in the individual difficulties, sensitivities, and performances (measured as the standard deviation of the individual posterior PDFs; blue histograms) with the variation in those quantities (measured as the standard deviation of the means of the individual posterior PDFs; red vertical line). It is clear that the difficulties can be very well determined and that the variation in difficulties is strongly dominated by the variation between the courses and not the variation within the courses. The same is true, although to a lesser extent, for the sensitivities of the courses. When looking at the performance of students, we see that there is a small population of students that have a very well defined performance, but that for a majority of students, the overall variation in fitted performances is only marginally larger than the uncertainty in the determination of individual performances. We conclude that course difficulty and course sensitivity are well determined and can be used at face value, but that the individual student performance needs to be taken with a grain of salt, and that inferences are likely to be only practically applicable on a cohort basis, or at least when

aggregating a group of students in some meaningful way.

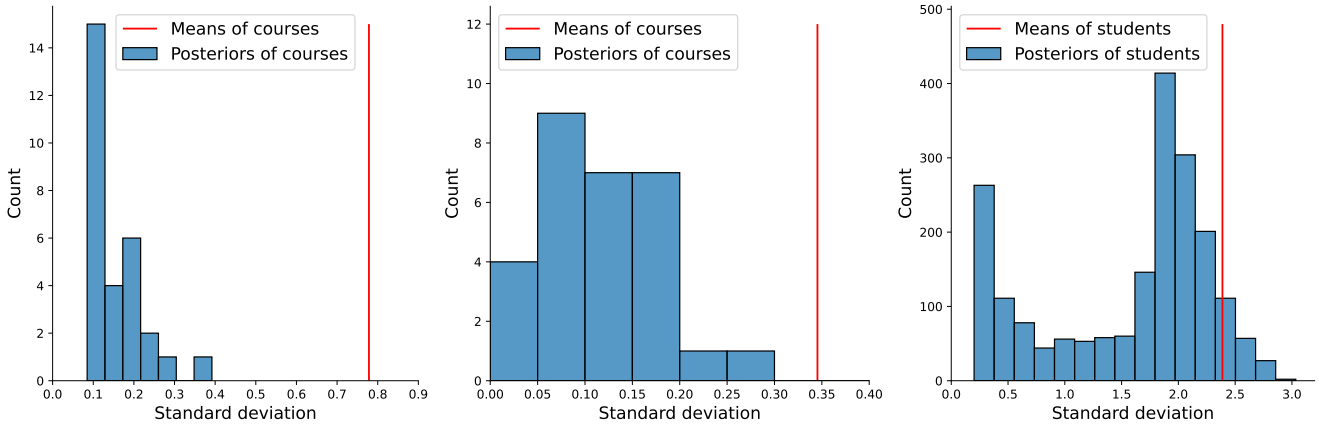

Figure 8. A comparison of standard deviations of course difficulty (*left*), course sensitivity (*middle*), and student performance (*right*). We compare the distribution of the standard deviations of the posterior PDFs of all individual courses (or students) in the histograms with the standard deviation of the means of those posteriors (vertical red line). For courses, the spread in difficulty and sensitivity between courses is much bigger than the uncertainty on the individual difficulties. For students, many posterior PDFs are much narrower than the spread between the means, but the difference is much smaller and the standard deviation of the mean is at the 93rd percentile of the distribution of individual standard deviations.

The recovered difficulties vary from 2 to 5.5, and the sensitivities from 0.4 to 1.6 and, with that, stay in a comfortable range given the model set-up. This illustrates that in a well-connected network the modelling as we describe in this paper works very well and results in well-defined and meaningful inferences about course properties.

5. Applications and Future Work

When enough confidence in the model exists, judging by arguments similar to those above, the application of models like these awaits. As we have said before, we strongly urge against the use of these models for direct inference of student performance. On the course and curriculum level, though, options are wider. Below, we outline several noteworthy application areas and potential directions for model expansion. The application areas are only maximally useful if the LA cycle can be closed (Clow, 2012) by feeding back the modelling results to the actual learning process, which is what we describe for various applications in the subsections below.

5.1 Predictive Modelling of Student Cohorts

Even though performance on a per-student basis is not meaningful, the aggregate of these over a cohort can readily be used (the uncertainty for the cohort being roughly a factor $\sqrt{N}$ smaller than the per-student uncertainty on the performance). With data about results of a cohort in the past, as well as (model-inferred) data on course difficulty of courses to be taken in the future by said cohort, progress of the population can readily be modelled. Typically, this is done with predictive modelling techniques (Jeffreys, 2015; Daud et al., 2017; Cui et al., 2019; Raji et al., 2021), but these can be strict on the data: missing data are hard for ML models to deal with, requiring the use of data imputation methods. The framework presented here does not mind missing data and can therefore be readily employed for all use cases of predictive modelling alluded to in the introduction, even in the case of incomplete and uncertain data. It can evaluate abilities and future performance with the data at hand, and if lots of data are missing for a student or course it only means that inference and prediction will be more uncertain, resulting in a wider posterior PDF for the variable one is interested in. Simply sampling the posterior PDFs per student and aggregating them will result in a full posterior PDF for aggregate results.

Such predictions can help in evaluating the likely throughput of students in an educational program—the fractional dropout (again, *who* will drop out will be less well determined unless methods like those described in Böttcher and colleagues (2020) and Lacave and colleagues (2018) are used) as well as the total number of credit points obtained (in, e.g., the Netherlands this is a crucial number for budgetary reasons) can be predicted, including a realistic bandwidth of uncertainty. Finances, as well as scheduling of classes, and even the utilization of real estate, can be informed by expected future results of the current student cohort.

One thing that can be done with individual student inference is flagging outlier results. As long as this indicator is not used by itself, it is helpful to have sensitive indicators of potential fraud on, e.g., take-home exams. As will be shown in a follow-up

paper (Caprani et al., in prep.), using information from their previous course results, as well as in-course progress, it is possible to flag potential fraudsters using the modelling techniques we present with much higher sensitivity than what could be done without modelling course difficulty and student performance in the process (i.e., based on grade statistics alone). Such flags can be used by the teacher and teaching assistant as an indicator for the student lagging behind and potentially needing some extra assistance. Care should then be taken, of course, not to accuse the student, since results will not be accurate enough for that, but only to signal potential learning problems, to be used in accordance with other signals that teachers might get. Taken over several courses, when a student's results get flagged very often, the uncertainty that the results are actually suspicious goes down and a student advisor or academic counsellor could take appropriate action.

5.2 Course and Curriculum Evaluation

Other than looking at outliers among the students, one can focus on outliers among the courses. A well-designed curriculum has courses that show a common thread, with appropriate, but not huge, excursions in topic, methodology, and perhaps difficulty. Aggregating results from individual modelling as presented in this paper could lead to some of the quantified curriculum analytics measures, as proposed by Ochoa (2016), which can be used in the accreditation process for higher education. Such curriculum-level aggregate results can be useful for the curriculum committee or an academic affairs department that is aiming for a well-balanced educational program to identify courses that need some attention in their design or their place or timing within the curriculum.

On the level of courses, the methodology presented here also has a concrete application. In many cases, exam results are corrected upward when the overall performance of a class is below expectation. Results are corrected to some desired class average. This is justified if the results are low because of an exam that turned out to be more difficult than intended. In reality, it can be a result of a difficult exam, but also of a below-average class. After the modelling as presented in this paper, one can objectively distinguish between the two and inform justified corrections to the grades by whoever decides about these corrections.

Eventually, also including a time dependency, as in, e.g., F. Chen and Cui (2020), could lead to full modelling of the entire *process* of students following large educational programs (Munguia & Brennan, 2020). Time dependencies can be included in the models we describe above if for example the student performance can be a function of time, where the functional form can be dictated (e.g., linear) or can be a free, but smooth, function of time, with, e.g., Gaussian processes (Roberts et al., 2013).

Modelling the curriculum as a whole, including time dependency, can also be a basis for the evaluation of the reception of education after sudden changes in the modes of education offered, which was very common in the early phases of the COVID-19 pandemic (Camargo et al., 2020). Programs and their learning goals did not in principle change, but many institutions needed to promptly move from the classroom to much more virtual education. With a well-calibrated model of the curriculum in place, the influence of such shocks on study results can almost instantly be measured and acted upon. Such moves to online learning tools are in themselves a great way to enhance the variety of data used in LA (Celik et al., 2023).

Including time dependence will also be critical in a curriculum where the concept of mastery learning (e.g., Block, 1979; Slavin, 1987) is applied. In such cases, explicit dependence of follow-up courses on the previous results needs to be implemented. Bayesian generative models are easily made hierarchical, enabling the modeller to make the presence of follow-up courses in the data dependent on results in earlier parts of the curriculum.

5.3 Higher-Dimensional Student Performance and Course Difficulty: Modelling a Skill Set

As indicated throughout the paper, the cognitive process of learning through courses has been very strongly simplified into a one-dimensional approach here. Even though the results look promising and prove to be useful with these simplifications, the move to higher dimensions is a trivial expansion of these models, which can result in insight into the multi-faceted process of course and curriculum design.

Higher education is more than just learning one trick. Several skills of different natures are developed, and some students may have natural talent for one particular set of skills and others for other skill sets, which ties in with the LA work of Chou and colleagues (2017). In any educational program, a mix of skills is necessary for success. The framework described in this paper can easily accommodate such a higher-dimensional view of study success. Moreover, it will be able to do so in a data-driven way; i.e., we can let the data dictate how many dimensions (or skills, if you want) are needed to describe the data well.

If along all dimensions the expected grade for that dimension can be written, analogously to Equation 1, as

$$g_{\mathrm{dim}} = \frac{g_{\mathrm{max}} - g_{\mathrm{min}}}{1 + \exp(s \cdot (\delta_{\mathrm{dim}} - \alpha_{\mathrm{dim}}))} + g_{\mathrm{min}}, \tag{4}$$

then the total grade for a course can be written as a weighted average of these dimensional grades, where the weights are free

parameters that are a property of the course (to what extent the course "measures" student performance in this dimension),

$$g = \sum_{\text{dim}} f_{\text{dim}} \cdot g_{\text{dim}}, \tag{5}$$

if the weighting factors f_{dim} are defined to sum to 1, i.e., if it is defined as the fraction of that course's grade that is determined by that particular dimension. Going to a higher number of dimensions increases the number of parameters that will be optimized to fit the data. Therefore it is only natural that the data will be better described by a higher-dimensional model. Nevertheless, increasing the dimensionality too much won't help reduce the uncertainties, which gives us a handle on how many dimensions will best describe the data. The number of dimensions for which the expected grade is optimally close to the observed grades, with small scatter, is a data-informed determination of the dimensionality of the data. Such a data-driven approach has important advantages over approaches in which these dimensions are predefined.

This optimal dimensionality might say something about different skills that can be tested in courses and exams. *Skills* is a rather broad term, though, and it would be interesting to see what these skills actually appear to be, in particular for researchers interested in the learning process through academic programs. The dimensions might show strong correlations with the discipline of the educational program, the faculty it is given in, or something more related to the form of examination (multiple choice versus open questions, or exams versus essays or presentations) or the Dublin Descriptors (Gudeva et al., 2012). It would be very interesting indeed if student performance along the several dimensions might be an aid in recommending particular electives or minors to students by student advisors or even apps that help a student choose, which would broaden the scope of applicability of these models.

6. Conclusions

In this paper we propose a Bayesian LA model that uses only course grades to simultaneously determine the latent variables of course difficulties and student performance. Our generative modelling approach avoids the black-box nature of ML (for example) and also provides insight beyond that of purely descriptive statistical methods: we derive PDFs for latent variables of both courses and students that can be used in objective interventions in the learning and grading process, as described in Section 5. The model is extensible and provides insights useful for predicting student and cohort performances and for evaluating course and curriculum. We demonstrate its effectiveness and limitations using artificial and real-world data sets.

We conclude that in order for the models to be able to constrain the latent variables, a few requirements must be met:

1. There must be significant overlap of student populations between courses. As was clear in the Open University data set, too little overlap in student populations makes it very hard to constrain properties of courses with respect to one another. The idealized experiments show that under ideal circumstances, a 10% overlap is enough. In a realistic setting like at the University of Amsterdam's College of Law, there is plenty of overlap to model course difficulty, course sensitivity, and student performance.

2. There must be variance in the grades. When one course does not differentiate between different students by giving everybody the same grade (or no grade at all), the properties of this course cannot be constrained. Moreover, the results of other courses in the same model will be skewed.

3. There must be some correlation between the grades for the different courses for the overlapping student population in the one-dimensional modelling presented here. This is easy to understand: course difficulties have one value per course, so the grades of students within one course should correlate with the latent student performance. If that is true for all courses, then there must be some correlation between the grades for different courses as well. If this is not the case, one could go to higher-dimensional course difficulty and student performance.

When these conditions are met for a data set, the difficulty of courses is typically well constrained, which means that inferences about the courses can be used to assess the curriculum under evaluation for homogeneity and consistency. On various levels of detail, one could investigate a curriculum, a set of courses, or sub-groups of students. The code is released publicly, including examples of how to extend its functionality.

Acknowledgements

The authors would like to thank the anonymous reviewers for constructive comments that helped in improving the clarity of this paper. This project makes use of PYTHON (Van Rossum & Drake, 2009), NUMPY (Harris et al., 2020), PYMC (Salvatier et al., 2016), ARVIZ (Kumar et al., 2019), NETWORKX (Hagberg et al., 2008), MATPLOTLIB (Hunter, 2007), and SEABORN (Waskom, 2021). MRH thanks the members of the Data Science Center of the University of Amsterdam for many useful discussions, in particular Joris Huese for inspiration that started the project and Alan Berg for the pointer toward the OU data set.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors declared no financial support for the research, authorship, and/or publication of this article.

References

Archambault, I., Janosz, M., Fallu, J.-S., & Pagani, L. S. (2009). Student engagement and its relationship with early high school dropout. *Journal of Adolescence*, *32*(3), 651–670. https://doi.org/10.1016/j.adolescence.2008.06.007

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, *113*, 177–194. https://doi.org/https://doi.org/10.1016/j.compedu.2017.05.007

Block, J. H. (1979). Mastery learning: The current state of the craft. *Educational Leadership*, *37*, 114–117. https://api.semanticscholar.org/CorpusID:142390161

Böttcher, A., Thurner, V., & Häfner, T. (2020). Applying data analysis to identify early indicators for potential risk of dropout in CS students. In A. Cardoso, G. R. Alves, & M. T. Restivo (Eds.), *Proceedings of the 2020 IEEE Global Engineering Education Conference* (EDUCON 2020), 27–30 April 2020, Porto, Portugal (pp. 827–836). IEEE. https://doi.org/10.1109/EDUCON45650.2020.9125378

Camargo, C. P., Tempski, P. Z., Busnardo, F. F., de Arruda Martins, M., & Gemperli, R. (2020). Online learning and COVID-19: A meta-synthesis analysis. *Clinics*, *75*, e2286. https://doi.org/https://doi.org/10.6061/clinics/2020/e2286

Celik, I., Gedrimiene, E., Silvola, A., & Muukkonen, H. (2023). Response of learning analytics to the online education challenges during pandemic: Opportunities and key examples in higher education. *Policy Futures in Education*, *21*(4), 387–404. https://doi.org/10.1177/14782103221078401

Chen, B., & Poquet, O. (2022). Networks in learning analytics: Where theory, methodology, and practice intersect. *Journal of Learning Analytics*, *9*(1), 1–12. https://doi.org/10.18608/jla.2022.7697

Chen, F., & Cui, Y. (2020). Utilizing student time series behaviour in learning management systems for early prediction of course performance. *Journal of Learning Analytics*, *7*(2), 1–17. https://doi.org/10.18608/jla.2020.72.1

Chou, C.-Y., Tseng, S.-F., Chih, W.-C., Chen, Z.-H., Chao, P.-Y., Lai, K. R., Chan, C.-L., Yu, L.-C., & Lin, Y.-L. (2017). Open student models of core competencies at the curriculum level: Using learning analytics for student reflection. *IEEE Transactions on Emerging Topics in Computing*, *5*(1), 32–44. https://doi.org/10.1109/TETC.2015.2501805

Clow, D. (2012). The learning analytics cycle: Closing the loop effectively. In *Proceedings of the Second International Conference on Learning Analytics and Knowledge* (LAK 2012), 29 April–2 May 2012, Vancouver, British Columbia, Canada (pp. 134–138). ACM. https://doi.org/10.1145/2330601.2330636

Cui, Y., Chen, F., Shiri, A., & Fan, Y. (2019). Predictive analytic models of student success in higher education: A review of methodology. *Information and Learning Sciences*, *120*(3/4), 208–227. https://doi.org/10.1108/ILS-10-2018-0104

Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F., & Alowibdi, J. S. (2017). Predicting student performance using advanced learning analytics. In *Proceedings of the 26th International Conference on World Wide Web Companion* (WWW 2017 Companion), 3–7 April 2017, Perth, Australia (pp. 415–421). International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3041021.3054164

Dawson, S., Poquet, O., Colvin, C., Rogers, T., Pardo, A., & Gasevic, D. (2018). Rethinking learning analytics adoption through complexity leadership theory. In *Proceedings of the Eighth International Conference on Learning Analytics and Knowledge* (LAK 2018), 7–9 April 2018, Sydney, New South Wales, Australia (pp. 236–244). ACM. https://doi.org/10.1145/3170358.3170375

Di Pietro, L., Guglielmetti Mugion, R., Musella, F., Renzi, M. F., & Vicard, P. (2015). Reconciling internal and external performance in a holistic approach: A Bayesian network model in higher education. *Expert Systems with Applications*, *42*(5), 2691–2702. https://doi.org/https://doi.org/10.1016/j.eswa.2014.11.019

Diaconis, P. (2009). The Markov chain Monte Carlo revolution. *Bulletin of the American Mathematical Society*, *46*, 179–205. https://doi.org/10.1090/S0273-0979-08-01238-X

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists* (1st ed.). Psychology Press.

Ferguson, R., & Clow, D. (2017). Where is the evidence? A call to action for learning analytics. In *Proceedings of the Seventh International Conference on Learning Analytics and Knowledge* (LAK 2017), 13–17 March 2017, Vancouver, British Columbia, Canada (pp. 56–65). ACM. https://doi.org/10.1145/3027385.3027396

Gardner, J. P., & Brooks, C. (2018). Evaluating predictive models of student success: Closing the methodological gap. *Journal of Learning Analytics*, *5*(2), 105–125. https://doi.org/10.18608/jla.2018.52.7

Gudeva, L. K., Dimova, V., Daskalovska, N., & Trajkova, F. (2012). Designing descriptors of learning outcomes for higher education qualification. *Procedia—Social and Behavioral Sciences*, *46*, 1306–1311. https://doi.org/10.1016/j.sbspro.2012.05.292

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In G. Varoquaux, T. Vaught, & J. Millman (Eds.), *Proceedings of the Seventh Python in Science Conference* (SCYPY 2008), 19–24 August 2008, Pasadena, California, USA (pp. 11–15). https://conference.scipy.org/proceedings/SciPy2008/paper_2/full_text.pdf

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Herodotou, C., Rienties, B., Verdin, B., & Boroowa, A. (2019). Predictive learning analytics "at scale": Guidelines to successful implementation in higher education. *Journal of Learning Analytics*, *6*(1), 85–95. https://doi.org/10.18608/jla.2019.61.5

Hoffman, D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*, 1593–1623. https://jmlr.org/papers/volume15/hoffman14a/hoffman14a.pdf

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jeffreys, M. R. (2015). Jeffreys's Nursing Universal Retention and Success model: Overview and action ideas for optimizing outcomes A–Z. *Nurse Education Today*, *35*(3), 425–431. https://doi.org/https://doi.org/10.1016/j.nedt.2014.11.004

Knight, S., Buckingham Shum, S., & Littleton, K. (2013). Epistemology, pedagogy, assessment and learning analytics. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (LAK 2013), 8–13 April 2013, Leuven, Belgium (pp. 75–84). ACM. https://doi.org/10.1145/2460296.2460312

Kumar, R., Carroll, C., Hartikainen, A., & Martin, O. (2019). ArviZ a unified library for exploratory analysis of Bayesian models in Python. *Journal of Open Source Software*, *4*(33), 1143. https://doi.org/10.21105/joss.01143

Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open University Learning Analytics dataset. *Scientific Data*, *4*, 170171. https://doi.org/10.1038/sdata.2017.171

Lacave, C., Molina, A. I., & Cruz-Lemus, J. A. (2018). Learning analytics to identify dropout factors of computer science studies through Bayesian networks. *Behaviour & Information Technology*, *37*(10-11), 993–1007. https://doi.org/10.1080/0144929X.2018.1485053

Lang, C., Siemens, G., Wise, A., & Gasevic, D. (Eds.). (2017, May). *Handbook of learning analytics* (1st ed.). Society for Learning Analytics Research (SoLAR). https://doi.org/10.18608/hla17

Macfadyen, L., & Dawson, S. (2012). Numbers are not enough: Why e-learning analytics failed to inform an institutional strategic plan. *Journal of Educational Technology & Society*, *15*(3), 149–163. https://drive.google.com/file/d/1TTNkuJmWOYsB_np3Et7ozDuqlSCqWMrd/view

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). CRC Press. http://xcelab.net/rm/statistical-rethinking/

McEneaney, J., & Morsink, P. (2022). Curriculum modelling and learner simulation as a tool in curriculum (re)design. *Journal of Learning Analytics*, *9*(2), 161–178. https://doi.org/10.18608/jla.2022.7499

Munguia, P., & Brennan, A. (2020). Scaling the student journey from course-level information to program level progression and graduation: A model. *Journal of Learning Analytics*, *7*(2), 84–94. https://doi.org/10.18608/jla.2020.72.5

Neal, R. (1994). An improved acceptance procedure for the hybrid Monte Carlo algorithm. *Journal of Computational Physics*, *111*, 194–203. https://doi.org/10.1006/jcph.1994.1054

Nicholls, M. (2007). Assessing the progress and the underlying nature of the flows of doctoral and master degree candidates using absorbing Markov chains. *Higher Education*, *53*, 769–790. https://doi.org/10.1007/s10734-005-5275-x

Ochoa, X. (2016). Simple metrics for curricular analytics. In J. Greer, M. Molinaro, X. Ochoa, & T. McKay (Eds.), *Proceedings of the First Learning Analytics for Curriculum and Program Quality Improvement Workshop* (PCLA 2016), 25 April 2016, Edinburgh, UK (pp. 20–24). International Educational Data Mining Society. https://ceur-ws.org/Vol-1590/paper-04.pdf

Raji, M., Duggan, J., DeCotes, B., Huang, J., & Zanden, B. V. (2021). Modeling and visualizing student flow. *IEEE Transactions on Big Data*, *7*(3), 510–523. https://doi.org/10.1109/TBDATA.2018.2840986

Richardson, J. T. (2015). Coursework versus examinations in end-of-module assessment: A literature review. *Assessment & Evaluation in Higher Education*, *40*(3), 439–455. https://doi.org/10.1080/02602938.2014.919628

Rienties, B., Cross, S., & Zdrahal, Z. (2017). Implementing a learning analytics intervention and evaluation framework: What works? In B. Kei Daniel (Ed.), *Big data and learning analytics in higher education: Current theory and practice* (pp. 147–166). Springer International Publishing. https://doi.org/10.1007/978-3-319-06520-5_10

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N., & Aigrain, S. (2013). Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *371*(1984), 20110550. https://doi.org/10.1098/rsta.2011.0550

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in python using PyMC3. *PeerJ Computer Science*, *2*, e55. https://doi.org/10.7717/peerj-cs.55

Schmitz, M., Scheffel, M., Bemelmans, R., & Drachsler, H. (2022). FoLA2—A method for co-creating learning analytics–supported learning design. *Journal of Learning Analytics*, *9*(2), 265–281. https://doi.org/10.18608/jla.2022.7643

Sense, F., van der Velde, M., & van Rijn, H. (2021). Predicting university students' exam performance using a model-based adaptive fact-learning system. *Journal of Learning Analytics*, *8*(3), 155–169. https://doi.org/10.18608/jla.2021.6590

Shah, C., & Burke, G. (1999). An undergraduate student flow model: Australian higher education. *Higher Education*, *37*, 359–375. https://doi.org/10.1023/A:1003765222250

Slavin, R. E. (1987). Mastery learning reconsidered. *Review of Educational Research*, *57*(2), 175–213. https://doi.org/10.3102/00346543057002175

Strecht, P., Cruz, L., Soares, C., Mendes-Moreira, J., & Abreu, R. (2015). A comparative study of classification and regression algorithms for modelling students' academic performance. In O. Santos, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds.), *Proceedings of the Eighth International Conference on Educational Data Mining* (EDM 2015), 26–29 June 2015, Madrid, Spain (pp. 392–395). International Educational Data Mining Society. https://www.educationaldatamining.org/EDM2015/proceedings/short392-395.pdf

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wise, A., Cui, Y., & Jin, W. (2017). Honing in on social learning networks in MOOC forums: Examining critical network definition decisions. In *Proceedings of the Seventh International Conference on Learning Analytics and Knowledge* (LAK 2017), 13–17 March 2017, Vancouver, British Columbia, Canada (pp. 383–392). ACM. https://doi.org/10.1145/3027385.3027446