On the Faithfulness of Simulated Student Performance Data

Michel C. Desmarais and Ildiko Pelczer
Polytechnique Montréal
michel.desmarais@polymtl.ca, ildiko.pelczer@gmail.com

Abstract. The validation of models for skills assessment is often conducted by using simulated students because their skills mastery can be predefined. Student performance data is generated according to the predefined skills and models are trained over this data. The accuracy of model skill predictions can thereafter be verified by comparing the predefined skills with the predicted ones. We investigate the faithfulness of different methods for generating simulated data by comparing the predictive performance of a Bayesian student model over real vs. simulated data for which the parameters are set to reflect those of the real data as closely as possible. A similar performance suggests that the simulated data is more faithful to the real data than for a dissimilar performace. The results of our simulations show that the latent trait model (IRT) is a relatively good candidate to simulate student performance data, and that simple methods that solely replicate mean and standard deviation distributions can fail drastically to reflect the characteristics of real data.

1 Introduction

Student cognitive diagnosis is commonly defined as estimating the probability of mastery of a set of skills by a given student. However, skills mastery cannot be directly measured. Instead, it is measured by observing performance results over a task, such as the successes or failures over a set of question items or exercises.

How can the accuracy of a cognitive diagnosis model be validated without direct measures of skill mastery? There are at least three, non exclusive means around this issue :

1. *Obtain indirect and independent measures of skill mastery*. Many studies rely on an independent source to estimate skill mastery and match the model prediction with this independent source. For example, Vomlel [9] asked experts to determine if a student mastered a set of skill in fraction algebra based on their answers to a test. The test data was used for training a Bayesian Network model and the prediction of the model was matched against the experts' judgment.

2. *Match predictions over observed items only*. Another approach consists in using solely the predicted outcome of observable items that can be directly matched to real data. No attempt is made at estimating skill mastery, and instead the approach relies on the assumption that hidden skills are correctly assessed if observed performance is accurately predicted.

3. *Generate simulated data*. The approach we investigate here consists in generating student performance data according to a predefined model for which skill mastery is defined for each student. This approach is commonly used in psychometric research where latent response models are validated against simulated data (see for eg. [4]). The approach has also been used for cognitive modeling within a number of studies and over different models such as the DINA [1] and a the Bayesian Network approach [7], to name but a few examples.

The obvious advantage of having predefined skills with simulated data is, however, plagued by the issue that the underling skill model may not reflect the reality. The models can be over simplistic, or they can misrepresent the relationships between skills and performance, and among skills themselves.

We investigate this issue by using four models of skills to generate simulated student data. We look at how close are the performances of a student model trained over real and simulated data, while ensuring that the simulated data reflects as closely as possible the characteristics of the real data. The student model for this study is a Bayesian approach to cognitive modeling, POKS [5].

The first data generation model is one of the simplest possible and it serves as a baseline. The probability of *item outcome* (generally defined as a success or a failure to a test item question, or to an exercise) is a function of the expected values from marginal probabilities of item success rate and student scores. The second data generation model relies on a Q-Matrix that defines the links between items and skills. The matrix is used to assign skill outcome probabilities, from which a data sample can be generated. A third approach is based on a standard approach in Monte Carlo simulations in which sample data is generated by a technique that preserves the correlations among variables (among items in our case). The fourth approach is based on latent trait modeling (IRT—Item Response Theory) [2].

A number of studies on generating simulated student data have been conducted for the latent trait (IRT) approach [10][3][6], but they were all done within the IRT framework, using the same underlying latent trait models both for simulating the data and for measuring the predictive accuracy of the student model constructed from this data. On the contrary, the current study uses a Bayesian approach as the student model and a makes comparison of widely different approaches in addition to IRT.

We explain each of the simulated data generation approach in greater details below before moving to the experiments and the results.

2 Expected Outcome Based on Marginal Probabilities

The simplest model for generating simulated data is based on the expected item outcome according to marginal probabilities, as represented by the student general skill level and the item difficulty. This model presumes of no conceptual or skill structure behind the items set. Each item is considered independent of the other and the outcome solely depends on the item difficulty and the ability of the student.

Within this framework, the generation of sample test outcome can be conceptualized as a random sampling process using the expected probabilities. Assuming two vectors of probabilities: (1) S, that represents the skills mastery level of students, and (2), Q, that represents the (inverse) difficulty of items, then, the outer product of the two vectors is a matrix $\mathbf{X} = Q \times S$ where each element, m_{ij}, represents the expected probability of student i mastering item j. In the current study, we forced the sampling process to exactly replicate the distribution of scores, S, by sampling a pedefined number of successes for each examinee.

Since the probability of an item x_{ij} being considered a success is solely dependent on the marginal probabilities, Q_i and S_j, we will refer to this model as the Marginal Probabilities sampling.

3 Q-Matrix Sampling

The second model we explore is based on a Q-matrix [8] which defines the links between items and skills. For example, assuming we have I items and K skills, and a response matrix of N students, then the Q-matrix and the response matrix are defined as:

$$\mathbf{Q} = \begin{bmatrix} q_{1,1} & \cdots & q_{1,K} \\ \vdots & \ddots & \vdots \\ q_{I,1} & \cdots & q_{I,K} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,I} \\ \vdots & \ddots & \vdots \\ x_{N,1} & \cdots & x_{N,I} \end{bmatrix}$$

For example, if an item x_1 involves only skills k_2 and k_3, then $q_{1,2}$ and $q_{1,3}$ will be set to 1, and all other entries for that item, $q_{1,\bullet}$ will be set to 0.

The skill mastery of a set of students can be computed as the dot product of the two matrices: $\mathbf{X} \cdot \mathbf{Q}$.

The generation of sample data from this Q-matrix consists in defining the probability of an item outcome as a function of the level of mastery of the set of skills it involves. By defining skill mastery in the range [0,1] (for which case the Q-matrix corresponds to a *capability matrix* as defined in [1]), then, the probability of a successful outcome to an item x_i is defined as the smallest of the mastery value of each skill involved for x_i. This is a heuristic estimate that reflects the requirement that all skills must be involved in order to correctly answer x_i.

Akin to the process described for *marginal probability sampling*, we can ensure that the scores distribution perfectly matches the real by fixing the number of item successes per

examinee. Sampling thus proceeds in a similar manner to the *marginal probabilities sampling* model, with the difference that instead of marginal probabilities, the item probabilities are derived from concept mastery. In turn, concept mastery is derived, in our experiment, from the student concept mastery distribution of the sample data and the capability matrix.

4 Covariance Matrix

Another mean of generating simulated student performance data is based on the idea of preserving the covariance (correlation) among items. This method is commonly used in Monte Carlo simulations. In the context of student test data, the method would stipulate that question items are interrelated and that a representative sample of simulated test data preserves the structure of correlation among items. This assumption is not unreasonable as we would, for example, expect that items of similar difficulty and that draw from the same skill set to show correlated student response patterns.

The generation of sample data based on item covariance relies on the Cholesky decomposition of the item covariance matrix. Assuming **L** is the upper triangular matrix of the Cholesky decomposition of the item covariance matrix, a first step is to generate a sample of correlated variables as:

$$\mathbf{S} = \mathbf{NL}$$

where **N** is an $N \times I$ matrix (number of students by number of items) of normally distributed independent random values having a mean of 0 and a standard deviation of 1. The sample data **S** will be an $N \times I$ matrix for which the item covariance matrix will approach the real data item covariance. It will have an expected mean of 0. The second step is to fit the distribution of this data's item success rate to the real data by adding the vector of real data item means to each row of **S** and, finally, to transform values to binary item outcome, setting values above 0.5 to 1 and 0 otherwise.

5 Latent Trait Models (IRT)

The last method of generating simulated student performance data relies on Item Response Theory, also known as *latent trait modeling*. As mentioned above, some authors have studied the faithfulness of this approach to replicate real data [10][3][6]. We refer the reader to [3] for a more elaborate description of this approach[1]

We use a 2 parameter logistic IRT model for generating the simulated data. According to this model, the probability of a successful outcome by an examinee s to an item i is defined as:

$$P(X_i \mid \theta_s) = \frac{1}{1 + e^{-a_i(\theta_s - b_i)}}$$

[1]Available from the ERIC Web Portal `http://eric.ed.gov/` under ref. ED414297 (accessed April 23, 2010).

where θ_s is the student's ability level, and where a_i and b_i are respectively the discrimination and difficulty levels of item i. The values for these three variables are directly estimated from the real data sample and therefore it is possible to replicate simulated data that reflects the real data. Estimates of the discrimination parameter is obtained with the R `ltm` package[2] and values for item difficulty and examinee ability are directly obtained through the logit transformation of the item average success rate and examinee percentage score. We also limit discrimination to values to the interval [0,4] and difficulty values to [-4,4], as is commonly done for IRT with small samples.

6 Experiments

We mentioned in the introduction that the issue with simulated student performance data is to determine how far the simulated data is representative of the complexity of the real student performance. To address this question, we train the POKS student model [5] over real and simulated data sets and compare its predictive performance across each condition. The simulated data sets are generated to closely resemble the real data according to the underlying model. The four models described above are used for simulated data: (1) **MP sampling**, marginal probability sampling (section 2), (2) **QM sampling**, Q-matrix sampling (section 3), (3) **Covariance**, sampling based on preserving item covariance using the Cholesky decomposition (section 4), and finally (4) **IRT**, sampling based on the latent trait modeling (section 5).

6.1 Adaptive Testing Simulation

The results of the different simulated data models are compared in the context of simulated adaptive testing with the POKS model. The process of adaptive testing consists in choosing the most informative item to present to the student and to infer the outcome of other items based on the pattern of previous item outcomes.

The performance is measured as the percent-correct predicted item outcome. Items that have been asked represent observed evidence and are considered correct by definition. Thus, performance after all items have been observed always converges to 100%. At the beginning, when no items are observed, item outcome is based on average item success rate: if an item has a success rate above 50%, it is considered mastered, and not mastered otherwise. As new items are observed, the POKS model computes the probability of mastery of each item based on the pattern of previous item outcome, and the predictions are compared to the actual data to compute the percent correct performance.

In this experiment, a cross-validation process is used for the College mathematics data set and a leave-one-out process is used for the Unix data set because of the small number of records.

[2] cran.r-project.org/web/packages/ltm/ltm.pdf

6.2 Data Sets

The characteristics of the real data sets from which the simulated data is generated can be very influential in this investigation and therefore we provide some details about them here. The experiment is conducted over two data sets:

1. *Unix*. The *Unix* data set contains 34 questions items that have all been answered by 48 respondents. The average success rate is 53% and it contains a large array of skills and difficulty, with test scores varying from 1/48 to 45/48, and item success rate varying from 1/34 to 34/34.

   Skills decomposition of this data is done over 9 topics (”sys-admin”, ”awk”, ”basic” ”directories”, ”file permissions”, ”input-output redirection”, ”printing”, ”regular expressions” ”shell language”). These topics contain from 3 to 7 items and only one topic is associated with an item. In other words, the row sums of the Q-matrix is always 1.

2. *College Mathematics*. The *Math* data set is composed of 59 items, which were administered to 250 freshmen students at Polytechnique Montreal. Each item was analyzed by two domain experts who determined if it involved one of the following topics : (1) Algebra, (2) Geometry, (3) Trigonometry, (4) Matrices and Vectors, (5) Differential equations and (6) Integrals. Mean student score is 57%, ranging from 9/59 to 55/59.

   Contrary to the *Unix* data set, most items are linked from two to four topics (only 17 are single topic, 32 are linked to two topics, 9 to three topics, and 1 to four topics).

The simulated performance data is generated to reflect as closely as possible the characteristics of the two real data sets. The similarity of the simulated data can be compared to the real one by looking at the correlation between success rates of students and items. Table 1 reports a number of similarity measures that represent the averages for 10 simulated data sets (numbers in parenthesis represent the standard deviations):

- *Mean* and *Sim. mean*: The percentage of correct responses over the whole data set. This number is to be compared to 53% for *Unix* and 57% for *Math*. The data generation process for the *QM sampling* and *MP sampling* methods were devised to match exactly this parameter.

- *Cor. exami.*: Pearson correlation between the simulated and real respondent test scores.

- *Cor. items*: Pearson correlation between the simulated and real average item scores.

- *Cor. concepts*: Pearson correlation between the simulated and real average concept mastery scores of students. Concept mastery for the students is computed on the basis of the dot product $\mathbf{X} \cdot \mathbf{Q}$ (see section 3), but with a normalization that ensures the scores range is between [0,1]. This normalization corresponds to the notion of a *capability matrix* (see [1]).

Table 1: Similarity of simulated data with real data

Unix	Mean	Sim. mean	Cor. items	Cor. exami.	Cor. concepts	% diff.
QM	.53	.53 (.00)	.93 (.01)	.80 (.03)	.81 (.02)	26 (1)
MP	.53	.53 (.01)	.55 (.08)	.64 (.10)	.28 (.05)	43 (2)
IRT	.53	.57 (.01)	.98 (.00)	.98 (.01)	.88 (.00)	15 (1)
Covariance	.53	.53 (.04)	.97 (.01)	.03 (.13)	.40 (.06)	38 (1)
Math	Mean	Sim. mean	Cor. items	Cor. exami.	Cor. concepts	% diff
QM	.57	.57 (.00)	.84 (.01)	.03 (.05)	.57 (.02)	44 (0)
MP	.57	.57 (.00)	.55 (.04)	.78 (.04)	.20 (.03)	47 (0)
IRT	.57	.62 (.00)	.83 (.01)	.89 (.01)	.44 (.01)	40 (0)
Covariance	.57	.56 (.02)	.98 (.00)	.07 (.10)	.11 (.05)	42 (0)

- *% diff.*. Percentage of items with different outcome.

The patterns of similarity vary considerably across the different sampling methods, but the most consistent one is the IRT method, in particular for the *Unix* data set, with correlations of 0.98 for both item success rate and examinee scores. Whilst these correlations are very high, we find that 15% of items differ from the real to the generated samples. We will see from the data in table 2 that this 15% difference can considerably degrade the predictive performance if the items are chosen at random.

7 Results

The CAT simulations experiment results are reported in Figure 1. The graphs depict the predictive performance of POKS over the two data sets. The percent correct number of item outcome prediction (accuracy) is reported over the different experimental conditions. Both graphs start a 0% observations, where the accuracy corresponds to guesses based on item average success rate. They end at 100% of questions observed for each data set, where the accuracy converges to 1 because observed item outcome are considered correctly "predicted". For indicative purpose, a straight line is drawn that starts at the initial guess of the real data, $(0, y_0)$, and ends at (1,1). It corresponds to the theoretical baseline accuracy of random guesses over non-observed items and provides an idea of the prediction gain obtained with the student model (note that only the real data line is drawn). Standard errors over simulation runs are not shown on the graphs to avoid cluttering, but they are at most around 7% and have no significant affect on the general patterns observed.

The different curves correspond to the four methods respectively described in section 2 to section 5 (see section 6 for label correspondance).

Table 2 provides a single score for the predictive performance, termed here the *accuracy gain*. This score represents the gain from guessing the outcome based on the initial probabilities of items and its range is [0,1]. It provides a simple means of comparing the

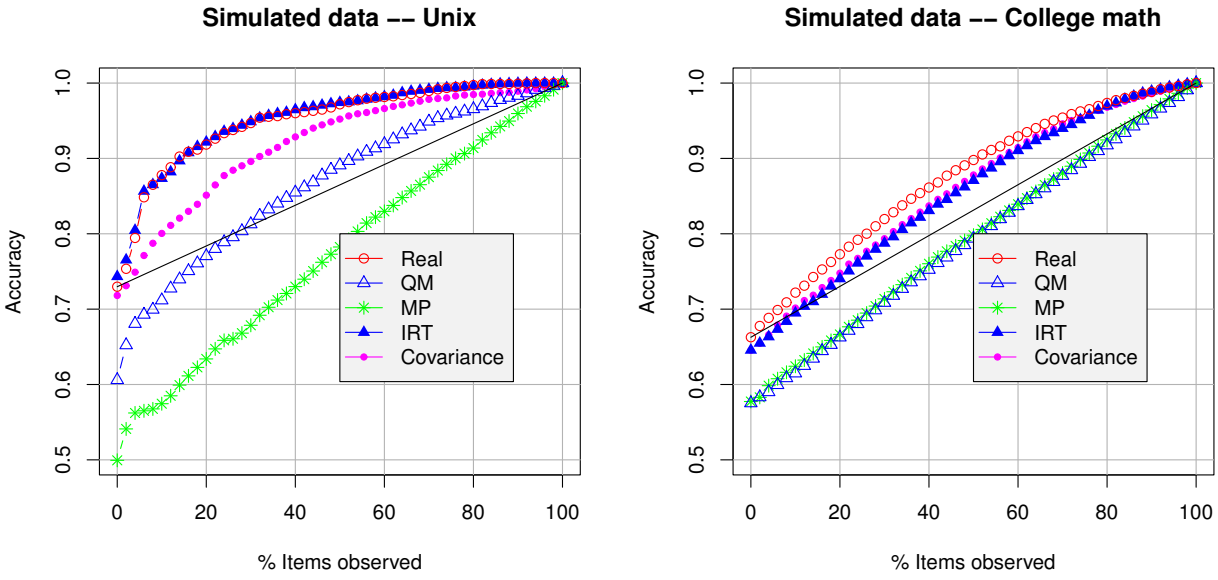

Figure 1: Results predictive accuracy simulation experiments with real student performance data compared with different models of simulated student data.

Table 2: Global accuracy gain over baseline

	Real	QM	MP	IRT	Covariance	15%
Unix	0.77 (0.0*)	0.43 (0.07)	0.14 (0.06)	0.80 (0.01)	0.58 (0.06)	0.29 (0.04)
Math	0.40 (0.02)	0.04 (0.01)	0.08 (0.01)	0.34 (0.01)	0.37 (0.01)	0.20 (0.03)

**Deterministic leave-one-out simulation*

overall predictive performances across the simulations and corresponds to the error reduction averaged over all intervals. It is computed as:

$$\text{accuracy gain} = \frac{1}{N}\sum_{i}^{N}\frac{y_i - \hat{y}_i}{1 - \hat{y}_i}$$

where N is the total number of intervals (we arbitrarily use 50), y_i is the accuracy at interval i (the x value) and $\hat{y}_i$ is the baseline accuracy at that same interval as represented by the straight diagonal of the figures (there exists one diagonal per curve but only the one for the real curve is represented in the figures).

For indicative purposes and in addition to the four methods, table 2 also reports a score corresponding to randomly changing the values of item outcome for 15% of the items, which is the proportion of items differing from the IRT simulated data to the real data.

In the case of the Unix data, the results indicate that the *IRT* method is able to generate data over which the POKS model is similar the performance, with accuracy gains of 0.77

for real data agains 0.80 for *IRT*. The *Covariance* method comes second with a performance of 0.58 instead of 0.77.

In the case of the Math data, the general predictive performance of all methods is substantially lower than for the Unix data. The *Covariance* and *IRT* methods both yield performance relatively close to the real data, but this time the *Covariance* method is closer to the real data performance.

8 Discussion

This investigation is limited to two real world data sets and to predictions based on a single student model, namely POKS. As such, further investigations are necessary to draw stronger conclusions. Nevertheless, we can still hint at some conclusions. First, the simpler methods of generating data, based on marginal probabilities and on concept mastery, yield simulated data that do not appropriately reflect the underlying structure of the real student performance data. However, the *IRT* method, based on the 2 parameter model (difficulty and discrimination), does appear to reflect the characteristics of real data, but not systematically for all data sets, as a non neglectible difference can be observed in the case of the Math data set. Furthermore, the *Covariance* method actually generates data for which the predictive accuracy is slightly closer to real data then the *IRT* method is. It also is close overall to the real data, standing at 0.37 accuracy gain compared to 0.40 for real data.

This investigation focused on models for generating data which allow their parameters to replicate real data characteristics, namely items difficulty, student skill levels, concept mastery as defined by the Q-matrix, and item covariance. Not all models allow this replication as readily as for these approaches. The DINA model used in [1] contains parameters that cannot be readily estimated from data, such as performance slips. Validating the faithfulness of such models is a desirable endeavour that would require means to estimate such parameters and constitutes an interesting research avenue. Indirectly, such investigations are in fact a means to validate if a model can actually reflect the characteristics of real data and, thus, they can be considered as an assessment of the external validity of a student model.

Turning back to the fundamental question of whether we can rely on simulated data to validate a student model, the simulations in this study suggest that simulated data from the 2 parameter *IRT* model can appropriately reflect some data set characteristics, but not with equal faithfullness for all data sets. It suggests that the validation of a model based on the indirect and independent measures of skill mastery may be indispensable to ensure a proper validation, as we outlined in the introduction. Alternatively, we could argue that the approach which consists in validating predictive performance over observable items only is just as indispensable. If we assume that the accuracy of a model for predicting item outcome is directly and monolitically linked to the accuracy of non observable parameters estimates of a model, then item outcome represents a good indirect measure of skills and concept mastery.

References

[1] Ayers, E., Nugent, R., and Dean, N. A comparison of student skill knowledge estimates. In *2nd International Conference on Educational Data mining, Cordoba, Spain* (2009), pp. 1–10.

[2] Baker, F. B., and Kim, S.-H. *Item Response Theory, Parameter Estimation Techniques*. Marcel Dekker Inc., New York, NY, 2004.

[3] Davey, T., Nering, M. L., and Thompson, T. Realistic simulation of item response data. Tech. rep., ACT Research Report Series 97-4, July 1997.

[4] de la Torre, J., and Douglas, J. Higher-order latent trait models for cognitive diagnosis. *Psychometrika 69* (September 2004), 333–353. 10.1007/BF02295640.

[5] Desmarais, M. C., Maluf, A., and Liu, J. User-expertise modeling with empirically derived probabilistic implication networks. *User Modeling and User-Adapted Interaction 5*, 3-4 (1996), 283–315.

[6] Harwell, M. R., Stone, C. A., Hsu, T.-C., and Kirisci, L. Monte carlo studies in item response theory. *Applied Psychological Measurement 20*, 2 (1996), 101–125.

[7] Millán, E., and Pérez-de-la-Cruz, J. L. A Bayesian diagnostic algorithm for student modeling and its evaluation. *User Modeling and User-Adapted Interaction 12*, 2–3 (2002), 281–330,.

[8] Tatsuoka, K. Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement 20* (1983), 345–354.

[9] Vomlel, J. Bayesian networks in educational testing. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems 12*, Supplementary Issue 1 (2004), 83–100.

[10] Yiand, Q., and Nering, M. L. Simulating nonmodel-fitting responses in a CAT environment. Tech. rep., ACT Research Report Series, December 1998.