applied sciences

MDPI

*Article*

# Probabilistic Models for Competence Assessment in Education

**Alejandra López de Aberasturi Gómez *, Jordi Sabater-Mir and Carles Sierra ***

Artificial Intelligence Research Institute (IIIA-CSIC), 08193 Barcelona, Spain; jsabater@iiia.csic.es
* Correspondence: alejandra@iiia.csic.es (A.L.d.A.G.); sierra@iiia.csic.es (C.S.)

**Abstract:** Probabilistic models of competence assessment join the benefits of automation with human judgment. We start this paper by replicating two preexisting probabilistic models of peer assessment ($PG_1$-bias and PAAS). Despite the use that both make of probability theory, the approach of these models is radically different. While $PG_1$-bias is purely Bayesian, PAAS models the evaluation process in a classroom as a multiagent system, where each actor relies on the judgment of others as long as their opinions coincide. To reconcile the benefits of Bayesian inference with the concept of trust posed in PAAS, we propose a third peer evaluation model that considers the correlations between any pair of peers who have evaluated someone in common: *PG*-bivariate. The rest of the paper is devoted to a comparison with synthetic data from these three models. We show that $PG_1$-bias produces predictions with lower root mean squared error (RMSE) than *PG*-bivariate. However, both models display similar behaviors when assessing how to choose the next assignment to be graded by a peer, with an "*RMSE* decreasing policy" reporting better results than a random policy. Fair comparisons among the three models show that $PG_1$-bias makes the lowest error in situations of scarce ground truths. Nevertheless, once nearly 20% of the teacher's assessments are introduced, PAAS sometimes exceeds the quality of $PG_1$-bias' predictions by following an entropy minimization heuristic. *PG*-bivariate, our new proposal to reconcile PAAS' trust-based approach with $PG_1$-bias' theoretical background, obtains a similar percentage of error values to those of the original models. Future work includes applying the models to real experimental data and exploring new heuristics to determine which teacher's grade should be obtained next to minimize the overall error.

**Keywords:** peer assessment; multiagent system; probabilistic model; comparative analysis; Bayesian network



**Citation:** López de Aberasturi Gómez, A.; Sabater-Mir, J.; Sierra, C. Probabilistic Models for Competence Assessment in Education. *Appl. Sci.* **2022**, *12*, 2368. https://doi.org/10.3390/app12052368

Academic Editors: Aida Valls and Agostino Forestiero

Received: 15 December 2021
Accepted: 19 February 2022
Published: 24 February 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



## 1. Introduction

Automated assessment and feedback of open-response assignments remain a challenge in computer science despite recent milestones in natural language processing. Competence assessment is a sensitive topic (current "AI Act" under discussion at the European Parliament pinpoints student assessment based on AI as high risk) with a clear impact on the certification of students as competent professionals and on their educational career progress.

Despite the efforts on opening the black box of neural networks, current neural models are rarely equipped with logical narratives of the decision chains that lead them to a final prediction or classification. Nevertheless, transparency and explainability are desirable requisites for automated assessment systems.

As a result, many researchers propose hybrid solutions combining the benefits of automation with human judgment. Specifically, they propose using peer assessments to help the teacher in the evaluation of students in large classrooms. Furthermore, numerous studies from the field of psychology point out that peer evaluation methods positively impact students' formative process, leading toself-reflection [1–3].

Among the hybrid solutions in the bibliography, there are many that make use of Bayesian models to infer the grade of an assignment given a list of peer assessments ([4–7]). On the other hand, Gutierrez et al. [8] successfully built a network of trust among peers

and the teacher that determines the relative importance that each peer's opinion has in the computation of the assignment's grade. This second approach, called PAAS, is very attractive as it exploits the naturalness with which a class of students evaluating each other can be modeled through multiagent theory.

The main objective of this article was to conceive a model that benefits from the Bayesian tradition in peer-assessment models while taking advantage of the concept of trust proposed by Gutierrez et al. [8]. As a result, we present *PG*-bivariate, a Bayesian model that translates the notion of trust among reviewers to a Bayesian approach thanks to its use of correlations among graders as the main feature to be learned by the model. Once *PG*-bivariate was implemented, we set out to compare it with one of the aforementioned Bayesian models and with PAAS in different experimental contexts. Given its theoretical robustness and the fact that it had been tested on an overwhelming number of more than 63,000 peers, the Bayesian model we chose to reimplement was $PG_1$-bias [4].

The three models (*PG*-bivariate, $PG_1$-bias and PAAS) were then compared in the context of competency assessment. All of them use a probabilistic approach to estimate a probability distribution for each automatic grade. Their inputs are always peer assessments and a given percentage of the teacher's grades (ground truths). Despite their commonalities, the relying idea varies for each model: whilst $PG_1$-bias [4] and *PG*-bivariate are Bayesian network models, PAAS [8] applies multiagent system theory. More specifically, it is a competency assessment model: Given a community of agents and a human leader whose assessing criterion is to be mimicked, PAAS builds a matrix representing each agent's trust in the rest.

The contributions of this work include (1) a Python reimplementation of PAAS, (2) a Python reimplementation of $PG_1$-bias, (3) a new model that integrates PAAS' use of trust measures with the Bayesian approach of *PG*-bivariate, and (4) a comparative analysis among the three models using simulated data and homogeneous units of measurement of the performance of the models.

This paper is structured as follows: In Section 1.2, we briefly review the concept of Bayesian network. After describing the materials and methods in Section 2, we present our models in Section 3. We experimentally evaluate the three models and compare them in Section 4. The discussion is presented in Section 4. Finally, we conclude in Section 5.

### 1.1. Related Work

The statistical models we present in this paper are part of a tradition of algorithms focused on score prediction. For instance, Bachrach et al. [6] makes use of a Bayesian model to grade tests. This model has two parts: a part that estimates the probability that a participant $p$ will know the correct answer to a question $q$, represented by a variable $c_{pq}$; and a second part that models the probability of each potential answer $r_{pq}$ of participant $p$ to question $q$ as a variable that depends on the correct answer to question $q$, $y_q$, and also on $c_{pq}$.

On the other hand, Sterbini and Temperini [5] exploited peer-evaluation among students to support the teacher when grading open-answers. To do so, they represented each student as a triplet of discrete variables: knowledge, judgment, and correctness, which are interrelated. Each student is asked to choose the best of three peer answers, and her choice is influenced by the triplet representing the student. The correctness of the three peer answers presented to her also makes an impact on that choice.

De Alfaro and Shavlovsky [9] created a web-based tool for collaborative grading and evaluation of homework assignments. Similarly to our setup, students played in this platform both the role of graders and gradees. In addition, to encourage quality peer reviews, they made the final grade received by the students dependent on (i) the consensus grade computed for the student's submission, (ii) an accuracy grade that measures the precision of the student in grading submissions, and (iii) a helpfulness grade that measures how helpful the reviews written by the student were. Such precision was computed

by comparing the grades assigned by the student with the grades given to the same submissions by other students.

Perhaps more related to Piech et al. [4]'s work, Ashley and Goldin [10] devised a hierarchical Bayesian model to mine peer assessment data. However, although they do account for graders' biases similarly to the models in this work, they mainly use the resulting posterior distribution over the conjoint parameter space to answer queries regarding the suitability of the employed rubric criteria.

### *1.2. Bayesian Networks*

Conditional probability distributions allow decoupling of joint probability distributions into sequences of conditional probability distributions over lower-dimensional spaces and marginal probability distributions. This decomposition eases the reasoning about the model and allows to better incorporate domain expertise. The methodology that constructs joint probability distributions by applying the chain rule to conditional and marginal probability distributions is known as generative modeling (Figure 1).

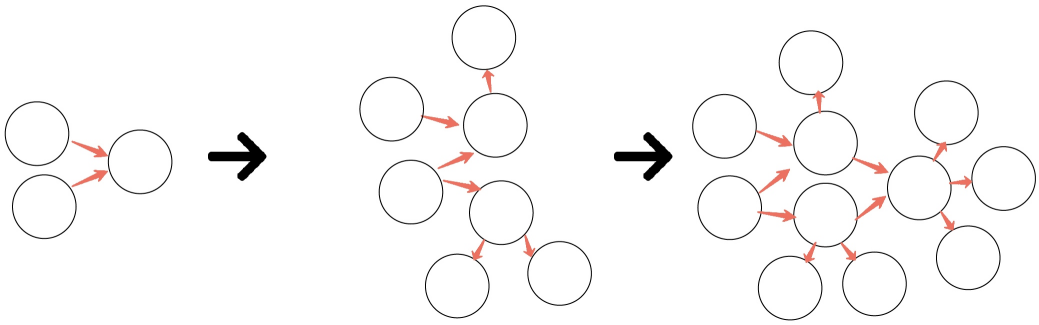

**Figure 1.** Iterative addition of variables to a generative model. This methodology constructs a joint probability distribution from intermediate conditional probability distributions.

Likewise, it is also possible to start with a coarse model represented by a joint probability distribution and then increase the complexity of the model by adding new variables as parents of that layer (Figure 2).

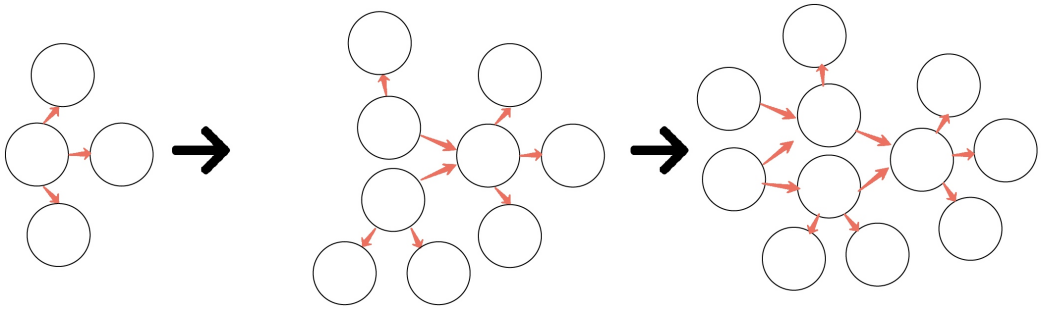

**Figure 2.** Reverse construction of a joint probability distribution. In this case, we begin with a marginal distribution over some few variables and then increase the complexity by conditioning this distribution over a new set of (ancestor) nodes.

A joint distribution that factorizes into marginal and conditional distributions can be represented by a probabilistic graphical model (PGM). In particular, this section will introduce the fundamentals of Bayesian networks, a kind of PGM that will be used during the description of two from three of the probabilistic models implemented for this research.

### *1.3. Representation*

A Bayesian network (BN) is a directed acyclic graph $G = (V, E)$ and a set of parameters $\Theta = \{\theta_1, \theta_2, ..., \theta_n\}$ labeling the probability distributions associated with the nodes in the graph. This graph represents a decoupled joint probability distribution, such that

1. The vertices represent the random variables that we model.
2. For each vertex $X_i$, there is a conditional probability distribution $\pi(X_i|\mathbf{pa}_i)$.

The chain rule for Bayesian networks states that a Bayesian network $M = (G, \Theta)$ can be expressed as the product of marginal and conditional probability distributions associated to its nodes:

$$\pi_M(\overrightarrow{X}) = \prod_{i=1}^{n} \pi(X_i|\mathbf{pa}_i; \Theta_i) \quad (1)$$

where $\overrightarrow{X} = \{X_1, X_2, ..., X_n\}$ and $\mathbf{pa}_i$ stands for *parent* nodes and denotes the set of variables with a direct edge towards $X_i$. The symbol $\Theta_i$ represents the parameters of the probability distribution associated with that same vertex. If a node has no parents, then the probability distribution associated with it is marginal. As a side note, when depicting a Bayesian network, observed nodes will be shaded.

In addition, some representations will introduce the plate notation, a method of representing variables that repeat in a graphical model. Instead of drawing each repeated variable individually, a plate (rectangle) is used to group variables into a subgraph that repeat together, and a quantity is drawn on the plate to represent the number of times the subgraph repeats in the plate. It is assumed that the subgraph is duplicated that many times, the variables in the subgraph are indexed by the repetition number, and any links that cross a plate boundary are replicated once for each subgraph repetition.

Before deepening into the dependencies between variables in a Bayesian network, let us introduce some important concepts that will be used in the presentation of the models implemented in this research, namely parents, children, ancestors, ancestral ordering, and Markov blanket.

- Given a node $X$ in a Bayesian network, its parent nodes are the set of nodes with a direct edge towards $X$.
- Given a node $X$ in a Bayesian network, its children nodes are the set of nodes with an incoming edge from $X$.
- Given a node $X$ in a Bayesian network, its ancestors are given by the set of all variables from which we can reach $X$ through a directed, arbitrarily long path.
- Given the set of all the variables modeled in a Bayesian network, $X = \{X_1, X_2, ..., X_N\}$, an ancestral ordering of the variables is followed when traversing the network; if every time we reach a variable $X$, we have already visited its ancestors.
- Given a node $X$ in a Bayesian network, its Markov blanket is given by its parents, its children, and the parents of its children.

### *1.4. Flow of Probabilistic Influence*

The structure of a Bayesian network contains information regarding how variables in the model interact with each other. In other words, it is possible to determine whether the injection of information about a variable $X$ updates our knowledge about another variable $Y$ in the graph.

Considering the case of only two variables $X$ and $Y$, if $X$ is a parent of $Y$, then any update in the probability distribution associated with $X$ will be reflected in changes in the probability distribution associated with $Y$. Likewise, knowing $Y$ will update our information about $X$.

If we now think about three variables $X, Y, W$, three main situations can be distinguished:

1. If $W$ is an intermediate node and all the edges go in the same direction (Figure 3), then an update in $X$ will be reflected in $Y$ if and only if $W$ is not an observed variable, and vice versa: an update in $Y$ will be reflected in $X$ if and only if $W$ is not an observed variable.
2. The same applies if $W$ is a parent of two children $X$ and $Y$ (Figure 4). Again, there will be a flow of probabilistic influence from $X$ to $Y$ if and only if $W$ is not observed.

3. Finally, if $X$ and $Y$ are parents of $W$ (v-structure, Figure 5), then the situation reverses, and there is a flow of probabilistic influence from $X$ to $Y$ if and only if $W$ is observed.

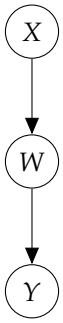


**Figure 3.** Case 1.

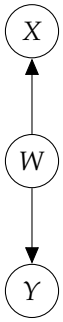


**Figure 4.** Case 2.

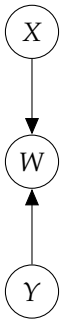


**Figure 5.** Case 3.

In short, it can be stated that in Bayesian networks, influence flow is stopped by observed nodes and nonobserved v-structures. A v-structure is observed if $W$ or any of its descendants is observed.

Formally, we say that there is a flow of probabilistic influence from $X$ to $Y$ if there is an active trail from $X$ to $Y$, where an active trail is defined as follows:

- Let $\mathcal{G}$ be a DAG.
- Let $X_1 \rightleftharpoons ... \rightleftharpoons X_2$ be a trail in $\mathcal{G}$.
- A trail is active given a set of observed variables $W$ if
  1. Whenever there is a v-structure $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, $X_i$ or one of its descendants is in $W$.
  2. no other node along the trail is in $W$.

Let $X$, $Y$, and $W$ be three disjoint sets of variables in $\mathcal{G}$. $W$ d-separates $X$ from $Y$ in $\mathcal{G}$ if $XY|W$ holds in $\mathcal{G}$. $XY|W$ holds in $\mathcal{G}$ if there is no active trail between any variable in $X$ and any variable in $Y$ given $W$.

Given a variable $X_j$ and its Markov blanket, $\mathbf{Mb}_j$, for any set of variables $\mathbf{X}_c \subseteq (\mathbf{V} \backslash \mathbf{Mb}_j)$:

$$\mathbf{X}_c X_j | \mathbf{Mb}_j \tag{2}$$

Hence, the Markov blanket of a variable in a BN d-separates that variable from the rest of the network.

## 2. Materials and Methods

All the analyses in this comparative study have been carried out with synthetic data. More specifically, for the generation of the data points injected into the Bayesian network models, ancestral sampling was used. According to this method, once a prior distribution $\pi_{\mathcal{S}}(\theta)$ and an observational model $\pi_{\mathcal{S}}(\tilde{y}|\theta)$ are specified, then we can iteratively generate an ensemble of reasonable model configurations and observations by sampling first from the prior

$$\tilde{\theta} \sim \pi_{\mathcal{S}}(\theta)$$

and then observations from the corresponding data generating process,

$$\tilde{y} \sim \pi_{\mathcal{S}}(\tilde{y}|\tilde{\theta})$$

Each simulated data point $\tilde{y}$ adds an independent sample to the synthetic database. This database is then used as input for the model to fit. Please refer to Section 3 for further information about the parameters of the Bayesian network models presented in this paper. With regards to the non-Bayesian model (PAAS), its equivalence with the original Java implementation in Gutierrez et al. [8] was tested in Section 3.2.2. As for the data fed to it, we replicated the synthetic experiment presented in Gutierrez et al. [8] for a class of 50 students.

## 3. Results

### 3.1. Probabilistic Models of Peer Assessment

In the last decade, the boom of massive online courses (MOOCs) has propitiated that some platforms that provide online educational resources direct efforts towards implementing technologies that somehow automate the assessment process. The goal is to help teachers evaluate usually large numbers of assignments in this teaching modality.

The variability of answers to open questions and their challenges to natural language processing (NLP) techniques make automatic assessment a limited tool to deal with this task. The more unique or creative an assignment is, the less appropriate it is to rely on purely computer-based assessment methodologies [11]. Many authors have proposed peer assessment techniques as a promising alternative to speed up evaluation in online courses. Moreover, according to these voices, this methodology may provide other potential benefits such as helping students see the task from an assessor's perspective and boosting self-reflection, as well as providing valuable feedback [3,12].

The three models of peer assessment that were studied in this paper are presented in this section. All of them are probabilistic and focus on an educational context and they all address a common question: How can we generate an automatic quality assessment of a submission not yet evaluated by the teacher in charge?

#### 3.1.1. Personalized Automated Assessments

Provided a community (a class) and a leader or special member from that community (the teacher), the aim of Gutierrez et al. [8] is to predict as accurately as possible the personalized assessments that a teacher in a class would make. The key idea behind this model is to unequally weigh the opinion of the members in the community (the students) so that the closer a student's marking style is to the teacher's style, the more relevant that student's opinion will be to predict the teacher's opinion.

The closeness in marking styles will be represented as a trust value. That is, the teacher's trust in a student depends on the similarity between the teacher's (past) assess-

ments and the student's (past) assessments of the same assignments. In what follows, a one-to-one mapping between assessed students (*a.k.a.* gradees) and assignments is presumed. In other words, for all the experiments in this work, every student completed one and only one assignment (*a.k.a.* exam).

In the case there were no commonly assessed exams by the teacher and a given student, the authors propose a key concept: indirect trust. In short, an indirect trust measure is computed as the reputation of the student within the community, where this reputation is biased towards the teacher's perspective.

Model

Let $\varepsilon$ represent a person who needs to assess a set $\mathcal{I}$ of objects and let $\mathcal{P}$ be a set of peers able to assess objects in $\mathcal{I}$. In the context of our analysis, this would be a teacher that wants to assess a number of assignments ($\mathcal{I}$) completed by students that are evaluated by other students ($\mathcal{P}$). Assessments made by peers $v \in \{\varepsilon\} \cup \mathcal{P}$ on an object $u \in \mathcal{I}$, noted $z_u^v$, are elements from an ordered evaluation space $\mathcal{E}$.

An automated assessment of $\varepsilon$'s opinion on an object $u$, noted $e_u^\varepsilon$, is a probability distribution $\mathbb{P} = \{x_1 \mapsto \alpha_1, x_2 \mapsto \alpha_2, ..., x_n \mapsto \alpha_n\}$, where $x_i \in \mathcal{E}$ and $\alpha_i \in [0, 1]$, with $\sum_i \alpha_i = 1$. A value $\alpha_i$ represents the probability that $\varepsilon$'s true assessment of $u$ is $x_i$. Hence, the more peaked the grading distribution on an object is, the more confident we will be that the automated assessment closely approaches $\varepsilon$'s. Inversely, the flat, equiprobable distribution will be the one denoting ignorance of $\varepsilon$'s true assessment.

Given a history of past peer assessments over $u$, $\mathcal{O}^u$, the ultimate goal of PAAS is to compute

$$\mathbb{P}(X_u^\varepsilon = x | \mathcal{O}^u)$$

That is, we aim to compute the probability distribution representing $\varepsilon$'s evaluation on every object given the assessments of peers on that object.

To that end, every individual evaluation $z_u^v$ from $\mathcal{O}^u$ is taken into account, and the probability $\mathbb{P}(X_u^\varepsilon = x | z_u^v)$ is computed as a function of the trust (expected similarity between previous assessments) that $\varepsilon$ has on $v$.

In our context, this trust is computed using the assignments graded in common by the teacher, $i$, and a student, $j$. In the case that there are no assignments in common, an indirect trust measure is obtained based on the notion of transitive trust: the trust that $i$ has on student $j$ can be computed from the trust $i$ has on student $k$ and the trust that student $k$ has on student $j$. As there are many possible paths connecting $i$ to $j$, appropriate aggregation functions have to be provided.

Direct Trust

When two agents $i$ and $j$ have commonly assessed one or more objects, a direct trust relationship between them $\mathbb{T}_{i,j}$ can be computed. This direct trust is modeled as a probability distribution on the difference between the evaluations performed by $i$ and $j$. This way, we keep information about whether $j$, for instance, underevaluates with respect to $i$, or overevaluates, or shows any other possible pattern. Information about their evaluation dissimilarities is summarized in that probability distribution.

The evaluation difference between two assessments performed by $i$ and $j$ is defined as

$$\mathit{diff}(i, j) = z_u^i - z_u^j$$

If $\mathit{diff}(i, j) > 0$, it follows from the above definition that agent $i$ over rates $u$ with respect to $j$. The opposite occurs if $\mathit{diff}(i, j) < 0$.

From the previous definitions, it is clear that a situation of complete agreement between two agents is represented by

$$\mathbb{T}_{i,j} = \{0 \mapsto 1\}$$

In what follows, we will denote that situation as $\mathbb{O}$.

Indirect Trust

If the leader has no objects assessed in common with a student $j$, a relation of indirect trust is computed. Since the model deals with probability distributions, it becomes necessary to define aggregation operators for probability distributions. These operators compute an indirect trust distribution from two trust distributions by combining differences in an additive way. The basic idea is that if the difference of opinions between $k$ and $j$ is $x$, and the difference of opinions between $k$ and the teacher is $y$, then the difference of opinions between $j$ and the teacher is $x + y$. Translating the above to probabilities and assuming independence of opinions, a combined distance distribution operator $\otimes$ can be defined as follows.

**Definition 1.** *Given trust distributions $\mathbb{P}$ and $\mathbb{Q}$ over the numeric interval $[-b, b]$ we define their combined distance distribution, noted $\mathbb{R} = \mathbb{P} \otimes \mathbb{Q}$, as:*

$$r(X = x) = \begin{cases} \sum\limits_{x_1 + x_2 = x} p(X = x_1) * q(X = x_2) & \text{if } x \in (-b, b) \\ \sum\limits_{x_1 + x_2 \leq -b} p(X = x_1) * q(X = x_2) & \text{if } x \leq -b \\ \sum\limits_{x_1 + x_2 \geq b} p(X = x_1) * q(X = x_2) & \text{if } x \geq b \end{cases} \tag{3}$$

This aggregation combines the distributions along a path between two peers.

From that definition, it follows that $\otimes$ is commutative, and its neutral element is the probability distribution representing the complete agreement between two peers ($\mathbb{O}$). In case there are several possible paths from the teacher to a student $j$, an aggregation operator $\oplus$ is defined to combine the aggregations computed along the different paths:

**Definition 2.** *Given probability distributions $\mathbb{P}$ and $\mathbb{Q}$ over the numeric interval $[a, b]$, we define $\mathbb{P} \oplus \mathbb{Q}$, as*

$$\mathbb{P} \oplus \mathbb{Q} = \arg \min_{\mathbb{T} \in \{\mathbb{P}, \mathbb{Q}\}} (EMD(\mathbb{T}, \mathbb{O})) \tag{4}$$

*with EMD standing for Earth mover's distance (a.k.a Wasserstein distance). (Informally, if the distributions are interpreted as two different ways of piling up a certain amount of dirt over the region D, the EMD is the minimum cost of turning one pile into the other, where the cost is assumed to be the amount of dirt moved times the distance by which it is moved [13]).*

This aggregation is optimistically performed; it assumes that the combined distance that is closer to $\mathbb{O}$ (i.e., the one that brings the student's and teacher's opinions closer) is the true one. This operator is both commutative and associative. Thus, the order in which distributions (i.e., paths) are combined is irrelevant.

Incremental Updates

Every time an object is assessed by a peer, PAAS launches an overall update of the trust values and hence of the student marks. This update can be simplified as follows:

1. Initially, the default direct trust distribution $T_{i,j}$ between any two peers $i$ and $j$ is the one describing ignorance (i.e., the flat equiprobable distribution $\mathbb{F}$). When $j$ evaluates an object $\alpha$ that was already assessed by $i$, $T_{i,j}$ is updated as follows:
2. Let $\mathbb{P}(X_u = x)$ for $x = diff(i, j)$ be the probability distribution of the assessment difference between $i$ and $j$. The new assessment must be reflected in a change in the probability distribution. In particular, $\mathbb{P}(X_u = x)$ is increased a fraction of the probability of $X$ not being equal to $x$:

$$\mathbb{P}(X_u = x) = \mathbb{P}(X_u = x) + \gamma \cdot (1 - \mathbb{P}(X_u = x)) \tag{5}$$

For instance, if the probability of $x$ is 0.6 and $\gamma$ is 0.1, then the new probability of $x$ becomes $0.6 + 0.1 \times (1 - 0.6) = 0.64$. As in the example, the value of $\gamma$ must be closer to 0 than to 1, for considerable changes can only be the result of information learned from the accumulation of many assessments.

3. The resulting $T_{i,j}$ is then normalized by computing the distribution that respects the new computed value and has a minimal relative entropy with the previous probability distributions:

$$\mathbb{T}_{i,j}(X) = \begin{array}{c}\text{argmin}_{\mathbb{P}'}\mathbb{P}'(X)\sum_{x'} p(X^{\alpha}x')\log\frac{p(X^{\alpha}x')}{p'(X^{\alpha}x')} \\ \text{such that} \quad \{p(X^{\alpha}x) = p'(X^{\alpha}x)\}\end{array} \tag{6}$$

where $p(X^{\alpha}x')$ is a probability value in the original distribution, $p'(X^{\alpha}x')$ is a probability value in the potential new distribution $\mathbb{P}'$, and $\{p(X^{\alpha}x) = p'(X^{\alpha}x)\}$ specifies the constraint that needs to be satisfied by the resulting distribution.
These direct trust distributions between peers are stored in a matrix $\mathcal{C}$.

4. To encode the decrease in the integrity of information with time *(information decays with time, I may be sure today about your high competence playing chess, but maybe in five years time I will be no longer sure if our interactions stop. You might have lost your abilities during that period)*, the direct trust distributions in $\mathcal{C}$ are decayed towards a decay limit distribution after a certain grace period. In our case, the limit distribution is the flat equiprobable $\mathbb{F}$. When a new evaluation updates a direct trust distribution $T_{i,j}$, $T_{i,j}$ is first decayed before it is modified.
5. The indirect trust distributions between $\epsilon$ and each peer are stored in a distributions vector $\overrightarrow{t_{\epsilon}}$.
Initially, $\overrightarrow{t_{\epsilon}}$ contains the probability distributions describing ignorance $\mathbb{F}$. When matrix $\mathcal{C}$ is updated, $\overrightarrow{t_{\epsilon}}$ is also updated as a product of its former version times matrix $\mathcal{C}$:

$$t_{\epsilon,j}^{k+1} = \bigoplus_{0<i\leq n} \mathbb{T}_{i,j} \otimes \mathbb{T}_{\epsilon,i}^{k} \tag{7}$$

6. If a direct trust distribution $T_{\varepsilon,j}$ exists between $\varepsilon$ and $j$, the indirect trust distribution $t_{\varepsilon,j}$ is overwritten with $T_{\varepsilon,j}$ after the update of the indirect trust distributions.

Please notice that the decay in the direct distributions $T_{i,j}$ affects the indirect distributions $t_{\varepsilon,j}$ as well. This is because the indirect trust distributions are computed as aggregations of combinations of direct trusts.

#### 3.1.2. Tuned Models of Peer Assessment in MOOCs

Looking for an answer to the same question posed by PAAS, the authors from *Tuned Models of Peer Assessment in MOOCs* [4] propose a Bayesian network model to represent the variables affecting the grade that a peer gives to another. Bayesian inference differentiates itself from other forms of inference by extending probability theory to the parameter space.

In this case, the space over which the probability is allocated is $Z \times \Theta$, where $Z$ is the observational space and $\Theta$ is the parameter space. Observations are denoted as $z_u^v \in Z$, where $u$ refers to the gradee and $v$ to the grader. On the other hand, $\Theta$ is the space parameter.

More specifically, $PG_1$-bias considers that $z_u^v$ is influenced by the following:

1. The assignment's true score, $s_u \in \mathbb{R}$. In the case of the implementation presented here, this is the teacher's grade.
2. The grader's bias, $b_v \in \mathbb{R}$. This bias reflects a grader's tendency to either inflate or deflate their assessment by a certain number of percentage points. The lower these biases, the more accurate the grades will be.
3. The grader's reliability, $\tau \in \mathbb{R}$, reflecting how close on average a grader's peer assessments tend to land near the corresponding assignment's true score after having corrected for bias. In this context, reliability is a synonym for precision or inverse

variance of a normal distribution. Notice that the reliability of every grader is fixed to be the same value.

The posterior probability distribution is computed as the product of the prior over the parameters times the likelihood function evaluated on the observations (times a constant depending only on the observations):

$$\pi_{\mathcal{S}}(s_u, b_v, \tau | z_u^v) \propto \int_{\Theta} \pi_{\mathcal{S}}(s_u, b_v, \tau) \pi_{\mathcal{S}}(z_u^v | s_u, b_v, \tau) \mathrm{d}\theta \tag{8}$$

The authors propose that the prior distribution $\pi_{\mathcal{S}}(s_u, b_v, \tau_v)$ can be decoupled into three marginal probability distributions:

$$\pi_{\mathcal{S}}(s_u, b_v, \tau) = \pi_{\mathcal{S}}(s_u) \pi_{\mathcal{S}}(b_v) \pi_{\mathcal{S}}(\tau) \tag{9}$$

That is, the parameters are mutually independent. These priors have the form, according to Piech et al. [4], of

$$\pi_{\mathcal{S}}(b_v) = \mathcal{N}(0, \frac{1}{\eta_0}) \tag{10}$$

$$\pi_{\mathcal{S}}(s_u) = \mathcal{N}(\mu_0, \frac{1}{\gamma_0}) \tag{11}$$

$$\pi_{\mathcal{S}}(\tau) = \mathcal{G}^{-1}(\alpha_0, \beta_0) \tag{12}$$

Whereas the likelihood function is given by

$$\pi_{\mathcal{S}}(z_u^v | s_u, b_v, \tau) = \mathcal{N}(s_u + b_v, \frac{1}{\tau}) \tag{13}$$

The resulting PGM is shown in Figure 6.

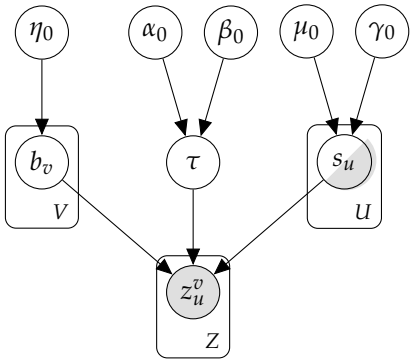


**Figure 6.** $PG_1$-bias [4].

Partially Known Parameters

As Figure 6 shows, once the evidence $z_u^v$ is injected, there is an active trail towards $s_v$, which implies that we do not need an example of a teacher's grade to make inference about how they would grade any student in the class. This active trail does not mean, however, that we are not allowed to introduce some ground truths grades $s_v$ in the Bayesian network. In that case, $s_v$ is said to be a partially known parameter, and its injection should result in a reduction of the variance of the posterior distribution. Half-shaded nodes represent partially known parameters in *BN*s.

3.1.3. *PG*-Bivariate: A Bayesian Model of Grading Similarity

We contribute in this section with a novel Bayesian model of peer assessment, *PG*-bivariate. The approach adopted here tries to reconcile the benefits of Bayesian inference with the concept of trust posed in PAAS [8]. The key idea is to use a probability distribution

to model the similarities as graders between any pair of peers in the system (teacher included). The chosen form for the probability density function modeling the similarities is a bivariate normal distribution. Figure 7 shows the graph representing the joint probability distribution of our model in a toy example with three students and a teacher, $\epsilon$. The small, dark squares represent the relationship variables $\mathcal{R}$ between graders.

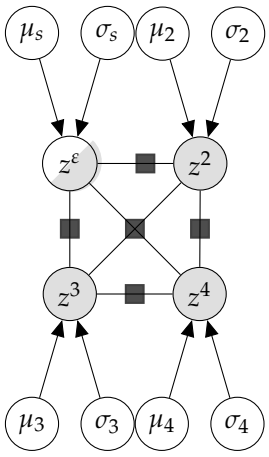


**Figure 7.** *PG*-bivariate.

More specifically, consider a class with $S$ students and a teacher $\varepsilon$. Each student $v_i$ assesses a number $m_i$ of peer students. The allocation of assignments to graders is such that every student is assessed by the same number $n$ of peers. Denoting the assessments by peer $v$ as $\vec{z^v}$, we wish to compute the grades that the teacher would give to each student, $\vec{z^\epsilon}$. The model proposes that $\vec{z^v}$ is sampled from a normal probability density function with parameters $\mu_v$ and $\sigma_v$. We could add complexity to the model and encode causal reasoning over these parameters by populating the layers above $\mu_v$ and $\sigma_v$. We limit, however, to this basic level of characterization, without adding any more variables influencing location or scale. Please notice that in this case, $z$ are vector variables containing the assessments of each peer to her assigned reviewees. In the case of $z^\epsilon$, it is a partially known vector: making it a completely observed variable would imply that the teacher has assessed all the students in the class, which is not the case.

In line with the concept of direct trust coined by PAAS' authors, we define a bivariate vector formed by the set of common assessments $\{(z_i^{v_1}, z_i^{v_2}), (z_j^{v_1}, z_j^{v_2}), \ldots, (z_k^{v_1}, z_k^{v_2})\}$ between two peers $v_1$ and $v_2$, hereafter denoted as $\mathcal{R}_{v_1,v_2}$. This sequence is sampled from a bivariate normal distribution such that

- The location vector, $\vec{\mu}_{v_1v_2} = (\mu_{v1}, \mu_{v2})$ is composed of each of the peers' location parameters separately.
- The covariance matrix $\Sigma_{v_1v_2}$ contains the individual variances in its diagonal. The off-diagonal components codify the correlations between $v_1$ and $v_2$ when grading.

Putting it all together, the equations for *PG*-bivariate model are

$$z^{v_i} \sim \mathcal{N}(\mu_{v_i}, \sigma_{v_i}) \quad \forall v_i \tag{14}$$

$$\mathcal{R}_{v_i,v_j} \sim \mathcal{N}(\vec{\mu}_{v_iv_j}, \Sigma_{v_iv_j}) \quad \forall v_i, v_j \tag{15}$$

where $\vec{\mu}_{v_iv_j} = (\mu_{v_i}, \mu_{v_j})$, $\Sigma_{v_iv_j} = \begin{pmatrix} \sigma^2_{v_i} & \sigma_{v_iv_j} \\ \sigma_{v_iv_j} & \sigma^2_{v_j} \end{pmatrix}$, and $(v_i, v_j)$ refers to a pair of graders having a set of evaluations in common.

We propose the following prior distributions for the hyperparameters:

$$\sigma_{v_i} \sim \text{Cauchy}(x_0, \gamma) \tag{16}$$

$$\mu_{vi} \sim \mathcal{N}(\beth, \daleth) \tag{17}$$

As for the modeling of the covariance matrix $\Sigma$, denoting by $D$ a diagonal matrix whose elements are the square root of the elements in the diagonal of $\Sigma$, $D = \sqrt{Diag(\Sigma)}$, then the correlation matrix $\Omega$ is related to the covariance matrix $\Sigma$ by

$$\Omega = D^{-1}\Sigma D^{-1} \tag{18}$$

Stan offers correlation matrix distributions, which have support on the Cholesky factors of correlation matrices. Cholesky's decomposition is unique: Given a correlation matrix $\Omega$ of dimension $K$, there is one, and only one, lower triangular matrix $L$ such that $LL^T = \Omega$. We call such a matrix its Cholesky factor, and, according to Stan, even though models are usually conceptualized in terms of correlation matrices, it is better to operationalize them in terms of their Cholesky factors.

Hence, we used the prior

$$L \sim \text{LkjCorr}(\eta)$$

The observations in this system are vectors of peer grades $\vec{z}^{v}$, whereas the inference target is the vector of teacher grades, $\vec{z}^{\varepsilon} = \vec{s}$. Please notice that in this case, no explicit mention is being made to the indirect trust. We limit our explicit modeling to the relationship between graders having a set of peers to assess in common. Nevertheless, from the graphical model, it can be seen that there exists a flow of probabilistic influence from any variable of type $\mathcal{R}$ to the rest of them. That is:, a change in the available information about the relationship between any pair of peers is reflected in an update of all the other variables $\mathcal{R}_{v_i v_j}$ for all pairs $\{v_i v_j\}$, including those containing the professor. In turn, this directs the flow of probabilistic influence towards the parameters of the distribution from which $\mathcal{R}_{v_i v_j}$ is sampled, namely $\vec{\mu}_{v_i v_j}$ and $\Sigma_{v_i v_j}$.

Partially Known Parameters

Similarly to $PG_1$-bias, *PG*-bivariate allows a flow of probabilistic influence from any peer's grades to the teacher's grading distribution without the need to have an example of how the teacher grades. However, it is possible to improve the inference results by introducing some of the teacher's grades in the system. Hence, the components in $\vec{s}$ will be considered partially known parameters as well.

### *3.2. Experiments*

The following results refer to synthetic databases, though future work includes applying the models to real data. Similar to in previous sections, the term *ground truth* is used here. In this context, ground truth is a synthetic data point with a known value that serves as input to the model.

#### 3.2.1. Experiments on Bayesian Networks

Posterior Predictive Sampling

Posterior predictive sampling (Equation (19)) is a useful method to compare predictions of new data based on past observations with the real data points. This comparison gives a sense of the estimations' quality and may help during the tuning of the hyperparameters.

$$\pi_{\mathcal{S}}(y|\tilde{y}) = \int \pi_{\mathcal{S}}(\theta|\tilde{y})\pi_{\mathcal{S}}(y|\theta)\mathrm{d}\theta \tag{19}$$

We sampled from the posterior predictive distribution of the two tested *BN* models, namely $PG_1$-bias and *PG*-bivariate, comparing the histograms of predictions with the histogram of samples. Sampling from the posterior predictive distribution helped us calibrate the models' hyperparameters for the prior distribution. In general, weakly informative and very weakly informative priors were proposed. However, we introduced some domain knowledge by centering the distributions of grades around positive numbers and restricting the variance parameters to only positive values, to name a couple of examples.

Recalling the prior distribution equation for model $PG_1$-bias (Equation (10)), the chosen values in our implementation for $(\alpha_0, \beta_0, \gamma_0, \eta_0, \mu_0)$ were $\alpha_0 = 0.01$ (very weak prior), $\beta_0 = 0.01$ (very weak prior), $\gamma_0 = 10.0$ (weak prior), $\eta_0 = 1.0$, and $\mu_0 = 5.0$, whereas in the case of $PG$-bivariate, the chosen hyperparameters were $x_0 = 0.0$ (weak prior), $\gamma = 0.25$ (weak prior), $\bar{\mu} = 5.0$, $\bar{\sigma} = 10.0$ (weak prior), and $\eta = 1.0$ (weak prior).

Figure 8 represents a histogram of predicted peer grades $z$'s from the posterior predictive distribution in model $PG_1$-bias for a class of 50 students (without ground truths). We observe the experimental histogram of observed peer grades in blue, $\tilde{z}$. The histogram of predictions ($z$) represents the distribution of data not part of the training set. This kind of prediction was repeatedly executed (20 times) to ensure that the results were consistent. Peer grades ranged from 0 to 10, following Equation (13).

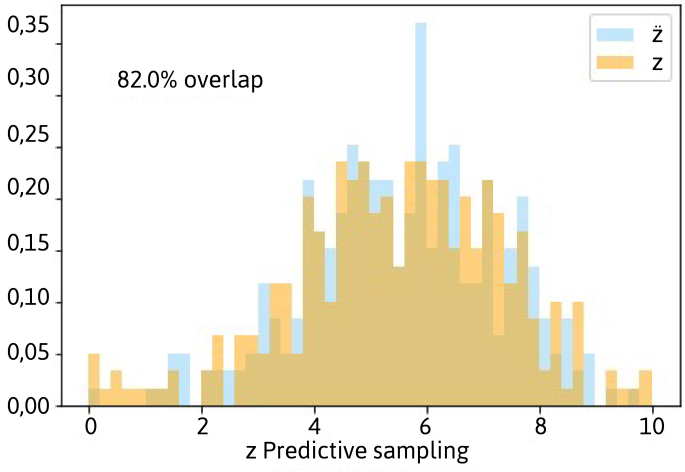


**Figure 8.** Posterior predictive sampling from $PG_1$-bias.

We calculated the percentage overlapping between the histogram of samples and the histogram of predictions using the vector of elementwise minima [14]. The overlapping range was between 0% (no overlap) and 100% (identical distributions). On average, we observed an overlapping of 82%. The range of values obtained for the overlapping in all the performed experiments was between 76% and 84%.

This value implies that the model captures the data generation system's dynamics, which results in samples $z$ that follow a very similar distribution to that from which $\tilde{z}$ were sampled.

A similar figure to Figure 8 is obtained when we perform a posterior predictive sampling of the model presented in this paper, $PG$-bivariate. The overlapping, in this case, fell within the range [73–81%], with an average value of 77% in 20 different runs (see Figure 9).

Although in this case, the model does not offer a causal explanation of the parameters $\mu_v$ and $\sigma_v$ from the normal distribution governing $z^v$ (Equation (14)), the dynamics of the data generation system are captured. Intuitively, one would have expected a notably worse performance than $PG_1$-bias for several reasons. First of all, $PG_1$-bias models each use peer grade as a random variable that follows its normal distribution. On the other hand, the definition of the location parameter as the sum of the real grade and the graders' bias allows a direct ascension of the flow of probabilistic influence towards $s_v$. In the case of $PG$-bivariate, we adopt the perspective of a passive observer, representing the set of grades emitted by a referee as a random variable following a normal probability density function. Notice that in this case, no mention of who is evaluated by whom is being made. Hence, the location parameter $\mu_v$ cannot be decoupled as in the former case to further ease the probabilistic influence flow.

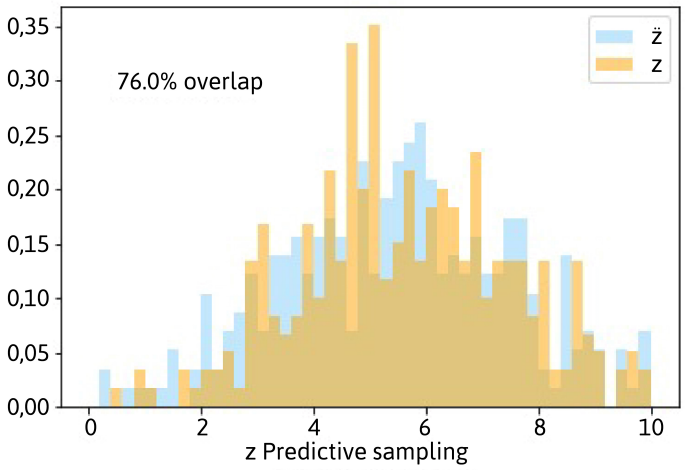


**Figure 9.** Posterior predictive sampling from *PG*-bivariate.

Studying the Error Evolution

To compare the performance of both models based on Bayesian networks, we studied the evolution of the models' root mean squared error (*RMSE*, Equation (20)) as a function of the number of observed true grades, keeping the number of students in a class, $S$, constant.

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - \tilde{y_n})^2} \tag{20}$$

where $\tilde{y}$ is the true value, $y$ is the prediction, and $N$ stands for the number of data points. Grades ranged from 0 to 10, both included. We examined the error evolution in both models as we introduced more teachers' grades as ground truths. We assessed two different criteria to determine which new grade to introduce in each step of the loop:

1. Random choice (baseline): The next observed ground truth is chosen randomly.
2. Total *RMSE* decreasing policy: At each iteration, we picked and observed the true grade (i.e., the teacher's grade) of that student whose assessment was introducing the highest root mean squared error.

As a result, we obtain four curves for each model:

- The red line shows the evolution of the estimations' *RMSE* as we introduce new ground truths following a random policy.
- The yellow, discontinuous line shows the evolution of the estimations' *RMSE* without considering the known ground truths to correct for overly optimistic low error values. Additionally, in this case, a random policy for ground truth injection is followed.
- The blue line shows the evolution of the estimations' *RMSE* as we introduce new ground truths following an *RMSE* decreasing policy.
- The violet, discontinuous line shows the same information as the yellow line for the case of the *RMSE* decreasing policy.

Figure 10 shows the resulting four lines in the case of $PG_1$-bias. Looking at the pair of continuous lines, we can see how *RMSE* falls as the number of known ground truths increases in all cases. As expected, computing the *RMSE* of all the grades (ground truths included) produces more optimistic errors. Moreover, it seems that the *RMSE* of the unknown grades (discontinuous lines) decreases much more slowly than for the continuous counterpart, that is, adding new ground truths does not reflect quickly on better quality of the predictions. According to Figure 6, introducing a ground truth $s_u$ has an impact on the bias $b_v$ of those peers $v$ who have made an assessment of $u$: $z_u^v$. There is also a flow of probabilistic influence towards the grader's reliability $\tau$, which in the case of *PG*1-bias is common to all students. The informativeness of the prior determines the relative relevance that observed data and domain expertise have on the posterior distribution parameters.

Hence, we believe that the choice of an informative prior over $b_v$ is preventing the model from showing an erratic behavior as new data is incorporated, but it also makes it more insensitive to the observations.

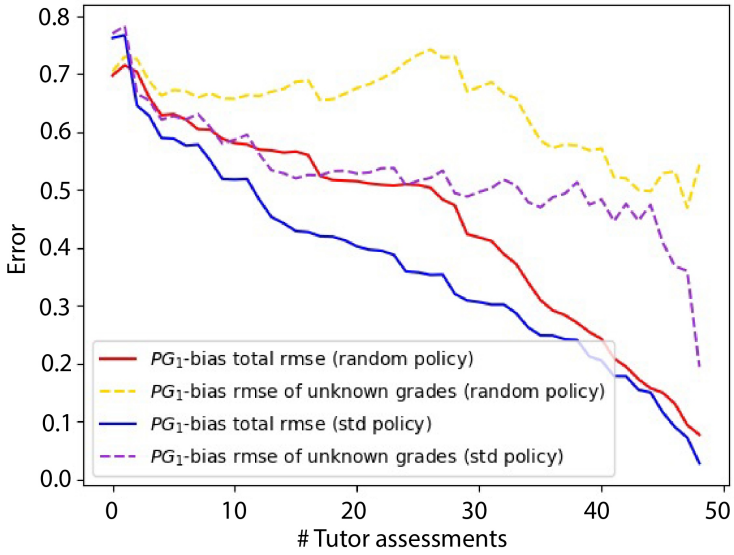


**Figure 10.** Reported *RMSE* for model $PG_1$-bias.

In general, the random policy yielded worse results for the discontinuous and continuous pairs of lines (red line showing higher error than the blue line, and yellow line showing higher error than the violet line).

We can see in Figure 11 the corresponding curves for *PG*-bivariate. In this case, the value of the *RMSE* was higher, which indicates that our predictions are, in general, worse than those of $PG_1$-bias. We see again that the random policy yields worse results than the *RMSE*-decreasing policy (red line above blue line, and yellow line above violet line). As in the previous case, the slopes of the continuous lines are sharper than those of the discontinuous lines, which implies that the probabilistic influence from the observed teacher grades towards the predictions does not flow easily. The discontinuous lines also make it evident that when the number of grades we are computing the *RMSE* with is small (last portion of the graph), fluctuations of the *RMSE* increase.

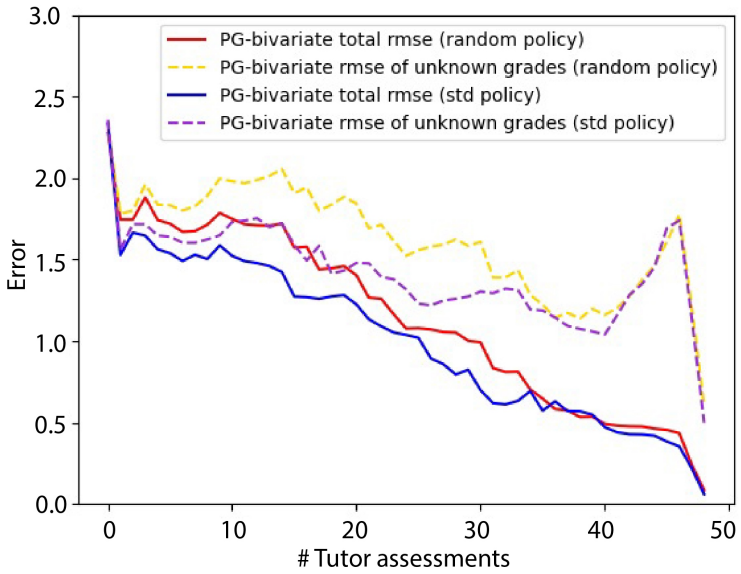


**Figure 11.** Reported *RMSE* for model *PG*-bivariate.

3.2.2. Experiments on PAAS

To make sure that PAAS software was being correctly implemented in Python, the synthetic experiment reported in Gutierrez et al. [8] was reproduced. We worked here under the premise that obtaining similar figures to the originals with our implementation would be a guarantee that the model functioning was being correctly replicated. This experiment consisted of a simulation of a classroom of 200 students with 200 submitted assignments, where each assignment was evaluated by 5 students (1000 peer assessments performed). In order to show a critical case, the authors simulated that half of the assignments were evaluated accurately by half of the students (that is, those students provided the same mark as the teacher), and the other half of the assignments were evaluated poorly (that is, randomly) by the rest of the class. In the simulation, they followed two policies to pick the next ground truth to observe as well. Such policies were a random one and another seeking to reduce the entropy of the grades by selecting the assignment with the highest entropy in its probability distribution, where entropy is computed as in information theory:

$$\mathbb{H}(\mathbb{P}(s_u|\{z_u^v\}_{v\in S}) = \sum_{\tilde{s}} \mathbb{P}(s_u = \tilde{s}|\{z_u^v\}_{v\in S}) \cdot \ln\mathbb{P}(s_u = \tilde{s}|\{z_u^v\}_{v\in S}) \quad (21)$$

The reported error line by Gutierrez et al. [8] is shown in Figure 12 (up). Down, we represent the obtained error for our implementation of PAAS in Python and a class of 50 students. We can see that although the original implementation shows a lower error for the first iterations, the behavior of both the original and our Python implementation is similar: both errors decrease as new ground truths are introduced in the system, and the entropy heuristic decreases the error faster than the random one. Furthermore, in both implementations, a crossing of the lines occurs when the number of observed ground truths is between 20% and 30%. From that point on, the performance of the heuristic-driven model is better.

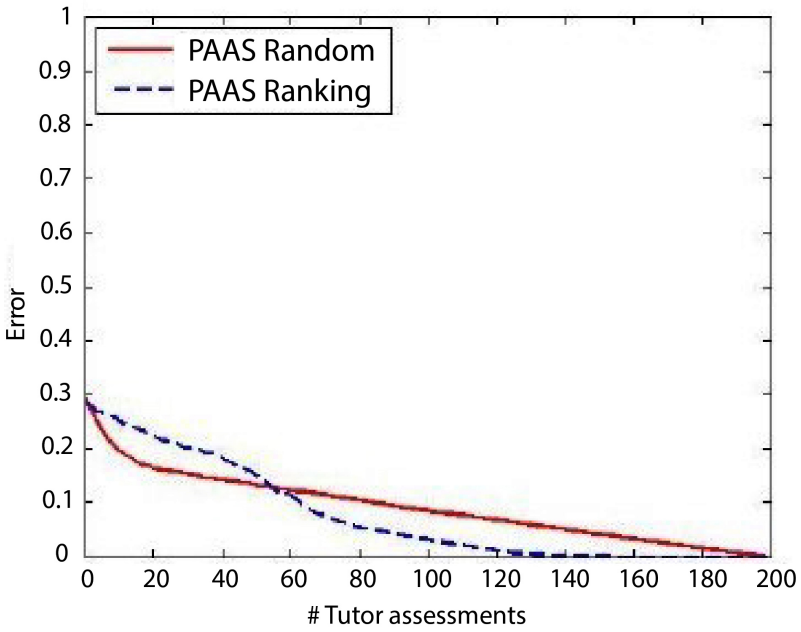


**Figure 12.** *Cont.*

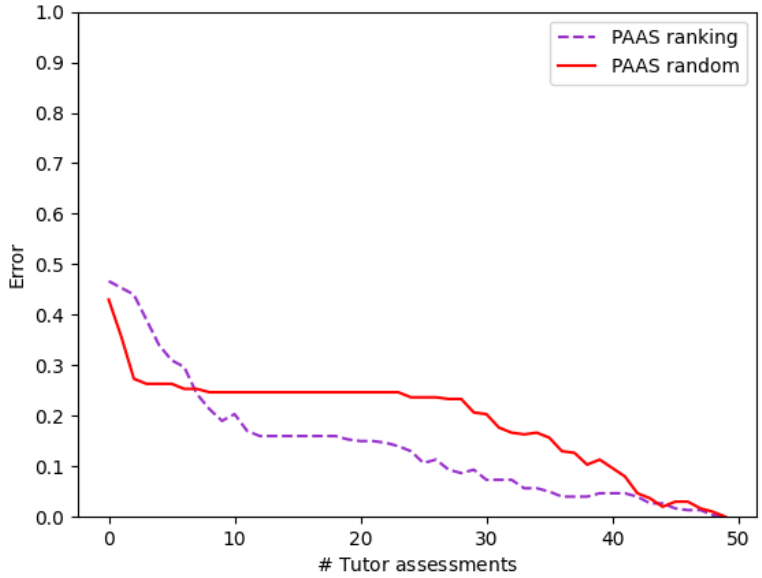


**Figure 12.** Percentage error using random (red line) and entropy decreasing (violet, discontinuous line) assessment order as a function of the number of observed teacher's grades for classes of 200 and 50 students using synthetic data. The **upper** figure represents the original implementation by Gutierrez et al. [8]. **Below**, our results are shown.

#### 3.2.3. Comparison of the Three Models

In order to compare the results, it was necessary that all of them were in the same unit. We adapted the $BN$ models and computed their percentage of error to make comparisons (Equation (22)). This is the error function that will be represented in the following figures:

$$\epsilon = \frac{\Sigma_{u \in \mathcal{S}} |\tilde{s}_u - s_u|}{S} \tag{22}$$

where $S$ is the number of students, $\tilde{s}_u$ is the $u$'s true grade, and $s_u$ is the prediction. The experimental setups on which PAAS was tested considered quality assessments (that is, the graders were asked up to what degree, in their opinion, did an assignment meet a set of criteria). Although the model allowed both qualitative (e.g., {bad, good, excellent}) or quantitative assessments, in the latter case, only integer grades were considered. Hence, for the comparisons to be fair, the Bayesian network models were adapted to predict integer grades ranging from 0 to 3 (both included). We ran this comparison mimicking the experimental arrangement described in the previous Section 3.2.2 for a synthetic class of 50 students.

Figure 13 shows the models' comparison using the best policy of each of them. From all of them, $PG_1$-bias keeps showing the lowest error when the number of known ground truths is small. Once past that region, both $PG_1$-bias and $PG$-bivariate show similar shapes downwards, although $PG_1$-bias keeps being solidly and continuously superior. Looking at PAAS' results and despite its initial results being poor, we can see a sharp decrease in error once some evaluations were introduced as knowledge in the network, showing, from that point on, a similar, though fluctuating, performance to $PG_1$-bias.

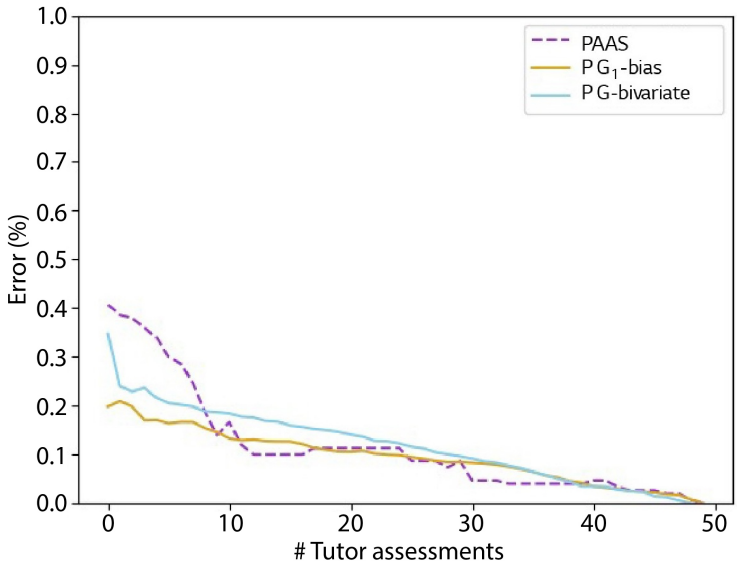


**Figure 13.** Percentage error as a function of the number of observed ground truth grades reported by the three studied models.

## 4. Discussion

The difficulties faced by neural models to make logical narratives of the decision chains that lead them to a final prediction pose a controversy on their applications in sensitive topics. One example is automatic grading and the ethical implications that it entails. These limitations have led the research to bet on hybrid solutions combining the benefits of automation with human judgment.

The first goal of this paper was to reproduce two models ($PG_1$-bias, by Piech et al. [4] and PAAS, by Gutierrez et al. [8]) of peer assessment. Both proposals estimate a probability distribution for each automatic grade using a probabilistic approach. These models' inputs are peer assessments and a percentage of the teacher's grades (ground truths). However, their theoretical foundations are different, and a common experimental setting was designed to make fair comparisons. All the code was written in Python, which implied the conversion of PAAS' original code from Java and the familiarization with Stan's Python interface (pystan) to reimplement $PG_1$-bias.

Our second goal was to implement a model that combined the first's powerful theoretical background with the second's multiagent-based ideas. As a result, *PG*-bivariate translates the notion of trust among reviewers to a Bayesian approach thanks to its use of correlations among graders as the main feature to be learned by the model. This allows for the obtention of automatic grades of the same quality as those of $PG_1$-bias without the need for complex causal reasoning over the distribution parameters representing each grader's grading style.

Despite the similarities between previous literature and the algorithms presented here, some important distinctions should be noted: first, in contrast to the Bayesian model by Bachrach et al. [6], the variables considered in $PG_1$-bias and *PG*-bivariate to predict the scores of the submitters are not only related to circumstances entailing the submitters, but also to the grades given by their peers and to the grading ability of such peers. Plus, instead of relying on highly abstract variables such as Sterbini and Temperini [5] do (e.g., judgment, knowledge), $PG_1$-bias and *PG*-bivariate work with more traceable variables such as true scores and grader's accuracy and reliability when correcting. Finally, and perhaps most importantly, all the algorithms in this work try to mimic the behavior that an expert would have had when faced with the task of grading a set of assignments. Hence, the trending opinion among peers is not what determines the quality of an assessment or what guides the prediction of an automatic score, in contrast to De Alfaro and Shavlovsky [9].

Regarding our experiments on Bayesian network models, the posterior predictive analysis showed that the distribution of new samples conditioned on the observed data was very similar to the distribution of samples. The posterior predictive distribution is influenced by the model configurations consistent with the relevant domain expertise and the observed data. This makes it ideal for informing predictions about future interactions with the latent system. Comparisons between the posterior predictive distribution and the observed data also measure how well our model approximates the latent system. Hence, we conclude that both models captured the data generating process.

When comparing these two models between them, specifically when analyzing the evolution of RMSE as a function of known ground truths (teacher's grades), we can see that $PG_1$-bias is slightly better. Not only does it report lower RMSE, but we observed that the computation times were shorter than in our model. According to the PyStan [15], most of the computation time during sampling with the NUTS algorithm is dedicated to calculating the Hamiltonian gradient. Looking at our computation times, *PG*-bivariate's phase space has a more problematic surface to the sampler than that one of $PG_1$-bias'. Hence, we conclude that although *PG*-bivariate uses a more simple definition of the distribution parameters (avoiding higher layers above $\mu_v$ and $\sigma_v$), in this case, a higher abstraction might be desirable to reduce the number of parameters to fit. Regarding the higher RMSE values reported by our model, our hypothesis is that better and more informed choices of the hyperparameters will yield lower errors in the future, further facilitating the flow of probabilistic influence between variables. However, it is worth noting that *PG*-bivariate showed a similar behavior to $PG_1$-bias in what policies are concerned: the random policy resulted in worse results than the one driven by the std error minimization policy. On the other hand, it is interesting to see that *PG*-bivariate offers remarkably good performance (the error of the predictions falls below 20% when only 20% of the ground truths are observed) despite its simplicity. The model can make predictions using the observed data and the correlations between graders, which somehow encode PAAS's direct trust. The flow of probabilistic influence between these *relationship* variables behaves in this case as PAAS' indirect trust. Given the success of PAAS in obtaining good automatic ratings, we find it interesting to continue exploring to what extent abstract variables can be dispensed within a Bayesian model that codifies the notion of trust between peers.

In Section 3.2.2, we showed that our Python implementation of PAAS, pyPAAS, reported a similar behavior to that by Gutierrez et al. [8]. To make fair comparisons among all models, we adapted the Bayesian inference models to compare with pyPAAS. More specifically, the synthetic data experiment was replicated for the three models. Under these conditions, $PG_1$-bias and *PG*-bivariate reported similarly low errors in situations of scarce ground truths. From 10% ground truths, the three models offer results of similar quality. It is worth mentioning that the calculation times for PAAS and our model were higher than those for $PG_1$-bias when the class size started to become relatively large.

Throughout this research, synthetic data was used. Future work will apply the models to real experimental settings. Exploring new heuristics to determine which ground truth to observe next is also an exciting research line to continue this work.

Finally, regarding the robustness of our method, the simultaneous concurrence of three circumstances guarantees that the model is shielded from potential malicious coalitions:

1. The small number (5) of peers assessing each assignment.
2. The fact that these graders are chosen randomly.
3. The fact that the process is entirely anonymous concerning the students (the graders do not know who are they assessing, and the gradees do not know the identity of their graders).

In future work, we plan to identify the scenarios where each of these models is preferable. Specifically, it would be interesting to confirm the hypothesis that the two models *PG*-bivariate and PAAS (that implicitly penalize bad raters) perform better than $PG_1$-bias under alliance and coalition dynamics among students. Applying the models to real experimental settings is also on the near horizon. Such settings may cover all those

situations that can be modeled as consisting of multiple agents with different reputations issuing an opinion (e.g., peer review). Further improvements of *PG*-bivariate should find a point of compromise between the conceptual simplicity of the model and the number of parameters that it requires in exchange. Finally, exploring new heuristics to determine which ground truth to observe next is also an exciting research line to continue this work.

## 5. Conclusions

Our paper compares two state-of-the-art automatic evaluation methods ([4,8]) with a new model. This new model combines the Bayesian background of Piech et al. [4]'s with the use of a trust graph over the referees proposed by Gutierrez et al. [8]. As a result, we obtain a hierarchical Bayesian model that dispenses with the choice of abstract variables in favor of others that are easily interpretable. Similar to PAAS, this modeling through a trust graph explicitly shields the algorithm against bad graders.

There remain several issues to be addressed in future work. First, it is necessary to find a point of compromise between the conceptual simplicity of the model and the number of parameters that it requires in exchange for said simplicity. For example, in the case of *PG*-bivariate, dispensing with variables above the parameters $\mu_s$ and $\sigma_s$ induces the necessity to calculate the parameters of all the bivariate distributions that describe correlations between students with some correction in common. This seems to cause a difficult phase space for the sampler because of the relatively long computation times. Second, we believe that further research that helps identify the scenarios where each of the compared models performs better is an interesting continuation line. For instance, it would be interesting to test whether the models that implicitly penalize bad raters perform better than $PG_1$-bias under alliance and coalition dynamics among students.

Similar to other studies of the area ([5,6,8]), we find that following a greedy policy that maximizes the entropy reduction induced by new observations yields better predictions of the grades. Exploring new heuristics to determine which ground truth to observe next is also an exciting research line to continue this work.

**Author Contributions:** Conceptualization, A.L.d.A.G., J.S.-M. and C.S.; methodology, A.L.d.A.G., J.S.-M. and C.S.; software, A.L.d.A.G.; validation, A.L.d.A.G., J.S.-M. and C.S.; formal analysis, A.L.d.A.G., J.S.-M. and C.S.; investigation, A.L.d.A.G.; resources, J.S.-M. and C.S.; writing—original draft preparation, A.L.d.A.G.; writing—review and editing, J.S.-M. and C.S.; visualization, J.S.-M.; supervision, J.S.-M. and C.S.; project administration, C.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research has been funded by ACCIO through the projects NanoMOOCs (COMRDI18-1-0010—RIS3CAT MEDIA), and ADDIA (ACE014/20/000039—INNOTEC) and by the CSIC through the project MARA (Intramural 202050E132). It has been funded also by the European Union Horizon 2020 FET Proactive projects "WeNet" (grant agreement Number 823783), "TAILOR" (grant agreement 952215) and "AI4EU" (grant agreement Number 825619).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The models presented in this study are openly available in GitHub at https://github.com/aloaberasturi/Probabilistic-Models-of-Competency-Assessment (accessed on 14 December 2021).

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| AI | Artificial intelligence |
| BN | Bayesian network |
| DAG | Directed acyclic graph |
| MOOC | Massive open online course |
| NLP | Natural language processing |
| PAAS | Personalized automated assessments |
| PGM | Probabilistic graphical model |
| RMSE | Root mean squared error |

## References


1. Schön, D.A. *The Design Studio: An Exploration of Its Traditions and Potentials*; International Specialized Book Service Incorporated: London, UK, 1985.
2. Tinapple, D.; Olson, L.; Sadauskas, J. CritViz: Web-based software supporting peer critique in large creative classrooms. *Bull. IEEE Tech. Comm. Learn. Technol.* **2013**, *15*, 29.
3. Kulkarni, C.; Wei, K.P.; Le, H.; Chia, D.; Papadopoulos, K.; Cheng, J.; Koller, D.; Klemmer, S.R. Peer and self assessment in massive online classes. *ACM Trans. Comput.-Hum. Interact. (TOCHI)* **2013**, *20*, 1–31. [CrossRef]
4. Piech, C.; Huang, J.; Chen, Z.; Do, C.; Ng, A.; Koller, D. Tuned models of peer assessment in MOOCs. *arXiv* **2013**, arXiv:1307.2579.
5. Sterbini, A.; Temperini, M. Correcting open-answer questionnaires through a Bayesian-network model of peer-based assessment. In Proceedings of the 2012 International Conference on Information Technology Based Higher Education and Training (ITHET), Istanbul, Turkey, 21–23 June 2012; pp. 1–6.
6. Bachrach, Y.; Graepel, T.; Minka, T.; Guiver, J. How to grade a test without knowing the answers—A Bayesian graphical model for adaptive crowdsourcing and aptitude testing. *arXiv* **2012**, arXiv:1206.6386.
7. Mi, F.; Yeung, D.Y. Probabilistic graphical models for boosting cardinal and ordinal peer grading in MOOCs. In Proceedings of the AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; Volume 29.
8. Gutierrez, P.; Osman, N.; Roig, C.; Sierra, C. Personalised Automated Assessments. In Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, Singapore, 9–13 May 2016; Jonker, C.M., Marsella, S., Thangarajah, J., Tuyls, K., Eds.; ACM: New York, NY, USA, 2016; pp. 1115–1123.
9. De Alfaro, L.; Shavlovsky, M. CrowdGrader: A tool for crowdsourcing the evaluation of homework assignments. In Proceedings of the 45th ACM Technical Symposium on Computer Science Education, Atlanta, GA, USA, 5–8 March 2014; pp. 415–420.
10. Ashley, K.; Goldin, I. Toward ai-enhanced computer-supported peer review in legal education. In *Legal Knowledge and Information Systems*; IOS Press: Amsterdam, The Netherlands, 2011; pp. 3–12.
11. Balfour, S.P. Assessing Writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review$^{TM}$. *Res. Pract. Assess.* **2013**, *8*, 40–48.
12. Admiraal, W.; Huisman, B.; Pilli, O. Assessment in Massive Open Online Courses. *Electron. J. E-Learn.* **2015**, *13*, 207–216.
13. *The Earth Mover's Distance (EMD)*; The Stanford University: Stanford, CA, USA, 1999.
14. Swain, M.J.; Ballard, D.H. Color indexing. *Int. J. Comput. Vis.* **1991**, *7*, 11–32. [CrossRef]
15. Stan Development Team. PyStan: The Python Interface to Stan. 2021. Available online: http://mc-stan.org/2 (accessed on 14 December 2021).