On the supervision of peer assessment tasks: an efficient instructor guidance technique

Jerónimo Hernández-González, and Pedro Javier Herrera

***Abstract*—In peer assessment, students assess a task done by their peers, provide feedback and usually a grade. The extent to which these peer grades can be used to formally grade the task is unclear, with doubts often arising regarding their validity. The instructor could supervise the peer assessments, but would not then benefit from workload reduction, one of the most appealing features of peer assessment for instructors.**

Our proposal uses a probabilistic model to estimate a grade for each test, accounting for the degree of precision and bias of grading peers. The grade that the instructor would assign to a test can help enhance the model. Our main hypothesis is that guiding the instructor through supervision of a peer-assessed task by pointing out to them which test to evaluate next can lead to improvement in the validity of the model-estimated grades at an early stage. Moreover, the instructor can decide how many tests to grade based on their own criteria of tolerable uncertainty, as measured by the model.

We validate the method using both synthetically generated data and real data collected in an actual class. Models that link the roles of the student as grading peer and as test-taker appear to better exploit available information, although simpler models are more appropriate in specific conditions. The best performing technique for guiding the instructor is that which selects the test with the highest expected entropy reduction. In general, empirical results are in line with the hypothesis of this study.

***Index Terms*—Peer assessment, Workload management, Probabilistic graphical models, Active machine learning**

I. Introduction

NEW methodologies have impacted all practices and structures of formal education. A well-studied evaluation methodology is *peer assessment*, where students evaluate each other's assignments (tests or any other activities) and provide feedback. Students need to be involved and take on a new role: they not only need to show their knowledge and skills in task solving, but for this new role, they also need to develop alternative skills to perform fair evaluations and provide constructive feedback. Peer assessment has been shown to promote learning and to have many positive effects on students [1]. However, its adoption is still limited [2].

As part of peer assessment, students may be required to provide a grade for the assignment. When peer-assessed grades are available, a dilemma [3] arises: can these grades be aggregated to calculate a final grade for the assignments? The aggregation is usually some simple kind of average. Instructors and students [3] have often questioned the validity of aggregated grades, which has restricted the use of peer assessment for evaluation. This, together with the extra workload required to manage the process, are common arguments explaining the limited adoption of peer assessment. In our work, we focus on peer assessment for grading and address these two main concerns simultaneously. Thus, the research question of this study is: Can a computational method be designed to allow instructors to supervise and weigh up peer-assessed grades with the objective of estimating valid final grades while at the same time minimizing their intervention?

We present a methodology that models the peer-assessment task, considering students' features and grades. Building on previous works [4], [5], we use probabilistic graphical models (PGM) [6], a type of model noteworthy for its interpretability. A Bayesian modeling enables (i) automatic peer-assessed grade aggregation for final grades, (ii) measuring the uncertainty associated with the model and the grades, and (iii) answering queries such as what can be done to reduce the uncertainty of the model. Identifying and requesting specific pieces of information to improve the model is the basic idea of active machine learning [7], which inspires this work. We aim to reduce the uncertainty of the peer-assessment model. Thus, we could use the model to identify the tests that involve the highest uncertainty. Our methodology selects tests one-by-one, requires the instructor to evaluate them and uses the instructor's grade to refit the model. Each model refit reduces the uncertainty and, eventually, will reach a value tolerable for the instructor. This assumes that the instructor provides the real grades and that the instructor's tolerable level of uncertainty can be reached (long) before they have supervised all the tests. If this is the case, our methodology potentially alleviates the workload of the instructor, which is convenient when the class is large or many peer-assessment tasks are carried out.

The main contribution of this work is a novel computational methodology that guides instructors in supervising peer-assessment tasks used for grading. Our proposal provides enhanced confidence in the aggregated grades while leaving the final decision of how many tests to review to the instructor. That is, it achieves a compromise between the reliance on peer-assessed grades and the effort that the instructor dedicates to supervising the task. To our knowledge, no similar methodology has been proposed before this in the related literature. We present a thorough analysis of the method and its various technical building blocks, namely:

- the model that describes the peer-assessment task; based on two state-of-the-art PGMs [5], we analyze up to five models with diverse underlying assumptions.
- the criterion for choosing the next test suggested to the

J. Hernández-González is a Serra Húnter Fellow at the Department Mathematics and Computer Science, University of Barcelona, Gran Via de les Corts Catalanes 585, 08007 Barcelona, Spain; e-mail: jeronimo.hernandez@ub.edu.

P.J. Herrera is with the Department of Software Engineering and Computer Systems, National University of Distance Education, Juan del Rosal 16, 28040 Madrid, Spain; e-mail: pjherrera@issi.uned.es.

Manuscript received December 31, 2021; revised –.

instructor to evaluate; we compare up to five selection criteria, from baselines that only use the available data, to complex model-based techniques.

Real data collected from an extended-response exam at the university educational level is used for validation. Synthetic data is also used to explore different realistic experimental scenarios and identify those where the proposed methodology is useful. Surveys of instructors' and students' views on the potential use of our method suggest plausible acceptance.

This document is structured as follows. First, we review the state of the art. Then, we present our methodology, including graphical models and selection techniques. Next, the experimental design is explained, and results are presented and discussed. To finish, we identify possible threats to the study's validity and suggest action plans, and draw some conclusions.

II. State of the art

A. Peer assessment

This study is concerned with peer assessment, defined by Topping [8, p. 250] as "an arrangement in which individuals consider the amount, level, value, worth, quality, or success of the products or outcomes of learning of peers of similar status". It is usually considered to be a technique for *formative assessment*, that is, a formative act where students (and instructors) identify the present knowledge and skills achieved [9], [10], which allows them to determine the way forward. Formative assessment requires the involvement of the students [9], [11], which is an essential element of peer assessment too. Similarly, Harlen and James [9] defined formative assessment in terms of *feedback*, which is the key strength of peer assessment according to Falchikov and Goldfinch [1]. The benefits for students of peer assessment have been well studied; it promotes learning and produces a positive impact on student training [1], actually exceeding the impact that might be expected from instructor assessment [8]. Other advantages include the opportunity to read their peers' solutions to the test and reflect upon them, the demystification of tests and evaluation, a boost in self consideration, as well as emotional benefits [12]–[16]. Most of the students find the feedback useful. Detractors request more specific, justified and constructive comments [17]. The benefits for the instructor include a rise in student engagement, as well as logistical benefits, principally the reduction of time spent in assessment [12]. However, researchers noted that preparation of the task and feedback support to students is key for the success of this methodology [18]–[20]. Among other things, students are usually provided with a rubric, i.e., a list of the criteria for assessing the test and degrees of achievement for each criterion, from poor to excellent [21]. The importance of a good rubric is studied in [22], proposing evaluation of the rubrics and reformulation of their high-variance items to improve peer-assessed grades. Peer assessment has been applied in all levels of formal education [23], from primary education to university [20], [24]–[28], and in informal education such as MOOCs [5], [22], [29]. It is usually combined with self-assessment [12], [22], where students assess themselves.

B. Modeling the assessment

The conventional practice of measuring the level reached by the student at a specific point in time for notification or teaching monitoring is known as *summative assessment* [9]. Despite its limited role in daily learning, it is fundamental in the overall education system. In this context, the level reached by students is usually recorded and reported as a numeric or ordinal grade. Abundant research in educational measurements has explored how to assign quantitative values to students based on data gathered in class: from analyses based on classical statistical theory to more or less complex models that aim to capture students' latent characteristics such as ability, progress or performance [30]. Several major statistical frameworks have been tested [30], such as generalizability theory, factor analysis, differential item functioning, cognitive diagnostic modeling, hidden Markov models, Bayesian knowledge tracing [31], structural equation modeling [32], latent class analysis [33], the popular item response theory [34], or Bayesian networks [35]. Many of these statistical frameworks are assimilable to each other; e.g., hidden Markov models can be seen as a type of Bayesian network (BN), which in turn are PGMs with a specific graph type and factorization [6]. These models are useful for assigning grades to the level reached by students which, in the context of summative assessment, are commonly used for notification. However, modeling can also provide actionable diagnostic information thus enabling formative assessment. Indeed, quantitative grades can be a valuable indicator of achieved skills, provided the grades and the associated feedback are made available promptly [35].

C. Modeling the assessment with PGMs

In this work, we focus on BNs or, in general, on PGMs. With a solid mathematical background, PGMs use an interpretable graph to encode the (conditional) dependencies between variables and determine a factorization of the joint distribution. Usually in educational research, the graph is drawn with expert domain knowledge and the model parameters are learned from data [35], although the structure can also be learned completely from data. Mislevy et al. [36] used BNs for managing uncertainty regarding students' knowledge and skills, exploiting the observable evidence about student behavior and task specification. Recently, they published a book on BNs for educational assessment [35] covering topics ranging from basic theory related to graphical models and Bayesian statistics, to model building and learning approaches. Another early BN-based modeling system was Andes [37], designed to help university students learn physics. It used BNs composed of domain-general and task-specific parts to perform students' knowledge tracing and predict their actions during problem-solving. Authors emphasized the ability of BNs to represent cognitive and pedagogical hypotheses for student modeling, although they also remarked on the need to test and refine these hypotheses through extensive empirical studies. To this end, many techniques are available for BNs checking or criticism: mutual information, model fit metrics, posterior predictive model checking, etc. [38], [39]; however, no consensus exists about the most appropriate procedure. In a recent review of BNs for educational assessment, Culbertson [38] categorized the state-of-the-art models into three degrees (Low/Medium/High) of detail in which students' cognitive processes are modeled. He identified open challenges like graph development (confronting discrete and continuous variables, identification of misspecification, etc.), parameter recovery (e.g., determining the amount of data/student numbers required for model calibration), or item selection. Our work addresses questions within at least the two latter topics.

Previous works usually take a Bayesian approach. Given initial beliefs (e.g., about students' knowledge), these can be updated with new data as it arrives [36]. Levy [40] provided a formal and high-level explanation of the differences between frequentist and Bayesian statistics, and their implications in the context of modeling in the educational domain. He emphasized that the Bayesian approach is particularly useful with small samples or when data arrives over time. Setting the model hyperparameters (initial beliefs) is commonly the primary concern. When no prior knowledge is available, so-called uninformative priors can be used. However, it has been reported that systems are not very sensitive to the specific values of the hyperparameters [37] and that collecting more data is the only way to improve the model.

D. Modeling peer assessment with PGMs

There are many BNs proposed in the literature for modeling the peer-assessment task, the complexity of which varies in order to capture more or fewer factors affecting the learning process. Piech et al. [5] proposed three models that capture the bias[1] and precision of each student, both as test-taker and as peer-assessor, coupling the performance of each individual within both roles (assuming that competent students are more precise as peer-assessors). Tests that are graded using an ordinal scale (discrete values with a given order between them) have been modeled too [41]. For multiple-choice tests, Bachrach et al. [4] proposed a model that accounts for student knowledge and item difficulty[2], and infers the probability that the student knows the answer to each question. Shalem et al. [43], considering interaction with multiple instructors, modeled the student's ability, the instructor's competence, and the question difficulty. Others extended the Piech et al. [5] key models to use assessor's deviation grades [44], ordinal grades [45] or understanding ordinal grades as censored data [46], as well as to account for other factors such as social interactions between students [47], cognitive diagnosis-derived student competency [48], or student effort expended [46]. We study the first and third models of [5] in the context of our methodology (explained in Sect. III-B1) and analyze their underlying hypothesis by contrasting them with simpler models. The second model in [5], a dynamic BN which considers a sequence of peer-assessment tasks, is omitted because analyzing the temporal dimension falls outside the scope of this study.

E. Dynamic graphical models

Dynamic BNs incorporate the temporal dimension in the model, accounting for the transition between time slices. Each time slice is a sub-graph that models a specific time. All time slices use the same variables and relationships between them, reducing the size of the network [37]. Sub-graphs are usually connected via edges among factors from adjacent time slices [38]. This type of modeling is known to have potential in ongoing assessment, and evidence shows that the validity of peer-assessed grades generally increases as the course progresses [22]. However, DBNs lack penetration in the field [49]. In fact, choosing a parametric form for students' knowledge as a function of time is currently understudied [38]. In his position paper, Reichenberg [49] provided a theoretical introduction to DBNs with practical examples. He identified several directions of future research on the use of DBNs for assessment. Many of them are related to the added complexity (compared to atemporal BNs) of modeling throughout time and challenging traditional hard assumptions in the transition between hidden states. Although DBNs are not considered in this study, our proposal applies to any BN (dynamic or not).

F. Crowdsourcing training data: an equivalent problem

Combining the contribution of many people to avoid individual bias and access social knowledge is known as *crowdsourcing* [50]. Due to conceptual similarities, many ideas that crowd-learning models incorporate also apply to peer-assessment modeling. In machine learning, crowdsourcing has been mainly used to collect labels for training data. Extensive literature exists [51]–[53] on aggregating a consensus label from the inaccurate annotations provided by multiple people. Founding studies modeled workers' reliability [52], [53], whereas others such as [54], [55] also considered instance difficulty, workers' competence or bias.

G. Active learning for assessment models

Modeling enables informed decision-making. When the answer or actions derived from these decisions can be used to improve the model, we are conducting active machine learning [7]. Active learning from crowds [56], [57] poses a dynamic scenario where decisions are made as the crowdsourcing task goes on, including deciding which part of the dataset requires more supervision [58], [59], who should be asked for a specific piece of information [60], etc. Items may be selected to minimize the expected posterior entropy, although this computation might be expensive in the case of continuous variables. It also suffers from the cold-start problem as little data is initially available for model learning [30]. In the educational domain, these ideas have been exploited for computerized adaptive testing when selecting an individualized set of elements appropriate for each test-taker [61]. It is claimed [61] that this approach better gathers knowledge about students' skills, leading to higher-validity estimations. BNs have been used in this context for a long time; e.g., Millán et al. [62] proposed a BN with a fixed structure composed of observed nodes (answers to given questions) and latent

[1] In a large study in MOOCs, Kulkarni et al. [22] found that, on average, peer-assessed grades are 7% higher than the ones assigned by instructors.

[2] de Alfaro and Shavlovsky [42] claim that errors are more due to hard-to-grade questions than to student imprecision.

variables representing the concepts of the subject at different granularity levels, as well as practical guidelines to set up the model parameters. In a class with multiple instructors, it has been used to match students and instructors so that students maximize their probability of passing a test [43]. Our focus is on assisting instructors; we should select the test with the real grade that would involve the maximum uncertainty reduction in the model if it was known.

III. Methodology

In this section, we describe the learning environment, i.e., peer assessment, and the procedure proposed to guide instructors in the supervision of the task.

A. *Learning environment*

This proposal complements peer assessment, where students engage in the evaluation of their classmates and themselves. Students are first test-takers, and then assessors or graders.

The usual dynamics consist of an activity or test completed independently (and, usually, individually) by students. Subsequently, students are given a (set of) tests carried out by their classmates to assess. That is, each student usually performs several assessments. This, in turn, implies that each student receives multiple assessments from different classmates. The instructor decides under which criteria the tests are assigned to students for peer assessment. It is common practice to provide instructions, for example, in the form of a rubric [21], to guide students in the evaluation task. With clear guidelines, students will read, understand and assess their classmates' answers to the test. Students are expected to make a *qualitative* assessment and provide valuable feedback on correct and need-improvement answers. Although initially peer assessment was not conceived for quantitative assessment or grading, instructions (using the very same rubric) can be provided to the student to transform their qualitative assessment into a grade.

We study the use of peer assessment for grading. Each student provides, apart from feedback, a single continuous grade per assessed test. The peer-assessed grades can be used to calculate a single final grade for each test. A simple approach is to assign the average value of the peer-assessed grades. However, lack of supervision, among other issues, raises many concerns in the teaching community. A different approach combines peer- and instructor-assessed grades, with the former representing a small percentage of the final grade. Nevertheless, in this case, the instructor evaluates all the tests, without any workload reduction. Our proposal opens an intermediate scenario enabling instructor workload lowering. With the model, our method aggregates the instructor- and peer-assessed grades and provides uncertainty estimates. It guides instructors on the selection of the tests to grade to ensure maximum uncertainty reduction in the model. If the instructor decides not to intervene, our method behaves similarly to a simple average of the peer-assessed grades. But the main benefit of our proposal is that it allows instructors to only grade a few tests until they reach an uncertainty estimation that they consider acceptable.

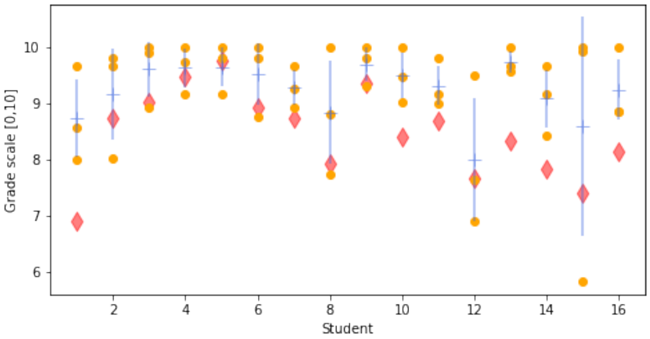


Fig. 1: Grades assigned by peers (orange dots), with mean value (blue crosses) and standard deviation (blue vertical lines), and instructor's grade (red diamonds) to the 16 students.

1) Participants and real data: For the present study, real data was collected from a class in a university master's degree. Sixteen students ($J = 16$) took an extended-response exam consisting of three questions to elaborate on theoretical-practical concepts of the subject's content. All the participants were informed of the use of the anonymized data in the present study and agreed to it. All the material used for preparing the peer-assessment task is available on the webpage associated with this study[3].

Three peers ($G = 3$) assessed each test. Each student carried out three peer assessments. The distribution of tests for peer assessment was randomized and checked to avoid closed groups. The peer-assessment task was carried out in a context of anonymity (double-blind). A rubric with the evaluation criteria was provided to students for standardization. The instructor used the same rubric to evaluate *all* the tests. Thus, assuming that instructors provide the real grades, ground truth grades are available for the analysis of our proposal. The collected grades are graphically displayed in Figure 1 where, for each test, the three peer-assessed grades and the instructor's grade are shown, with a measure of the mean and standard deviation of the peer-assessed grades. Figure 1 shows that peer-assessed grades tend to overestimate the instructor's grades (the numerical grades are available in the supplementary material). In some cases (e.g., students 1 or 13), all the peers consistently overestimated the grade by a large margin. In most of the cases, although overestimated, the real grade lies in one standard deviation from the mean of peer-assessed grades.

B. *Procedure*

The proposed methodology guides instructors through the supervision of a peer-assessed task to grade students' work. Without any intervention, the final grade is usually the average of the peer-assessed grades, which many instructors criticize due to lack of validity. If the instructor evaluates and grades all the tests, the peer-assessed grades make no contribution and there is no benefit for the instructor. This technique proposes an intermediate scenario: it provides estimates for the final

[3] https://jhernandezgonzalez.github.io/supp_pa_pgms.html

Algorithm 1 Pseudo-code for the proposed procedure.

```
Require: $J, M, \boldsymbol{z}$
Initialize model $M$ with peer-assessed grades, $z_u^v$
$\boldsymbol{S} := \{1, \ldots, J\}$    {Set of unsupervised tests.}
while instructor wants another one and not completed do
    Select next test to grade, $j \in \boldsymbol{S}$
    Instructor revises and assigns grade $s_j$ to that test
    With $s_j$, update model $M$ and model uncertainty $h$
    $\boldsymbol{S} := \boldsymbol{S} \backslash \{s_j\}$    {Remove j from unsupervised list.}
    print Current uncertainty $h$. Continue (y/n)?
end while
```

grades using a model that is enhanced with the instructor's grades as they review the tests. Model uncertainty measures are provided to help instructors decide how many tests to evaluate.

The procedure is applied when the peer-assessment task is completed and the instructor has not evaluated any test. It proposes a test to evaluate to the instructor, who provides a grade for it. The method will answer with an estimation of the remaining tests' grades, the model uncertainty, and, when asked by the instructor, a new test to evaluate. It is composed of a set of decisions and steps, namely:

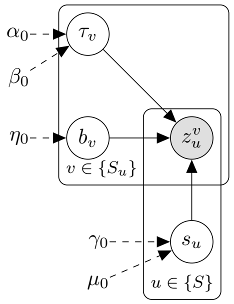
1) To choose (design) and assemble a probabilistic model of the peer assessment activity (see Section III-B1).
2) To decide how the next test that the instructor needs to supervise will be selected (see Section III-B3).
3) To update the model after each instructor evaluation and estimate the grades of those tests not yet evaluated by the instructor (see Section III-B2).
4) To stop when the instructor supervises all the tests or decides that the uncertainty involving the model's grade estimates is tolerable, using their own criteria.

After this, we provide empirically evidenced guidelines on decisions at 1 and 2. Algorithm 1 describes our methodology as a composite of these steps.

1) Probabilistic graphical models for peer assessment: As the instructor grades tests, the instructor's and peer-assessed grades can be compared and the abilities of the students as peer-assessors can be estimated. In this study, PGMs [6] are used to model students' abilities and performance, and the peer-assessment relationships between students. A PGM is a mathematical tool that allows for encoding conditional dependencies between random variables using a graph. A set of factors parameterizes the probability distribution. Specifically, we use Bayesian networks, a type of PGM with a directed acyclic graph and the conditional probability distributions of each variable given its parents in the graph as factors.

Among the different PGMs proposed for peer assessment in related literature, in this work we inspect two models proposed by [5] and test their underlying assumptions by comparing them with simpler models. The proposal of new PGMs for this task falls outside the scope of this work.

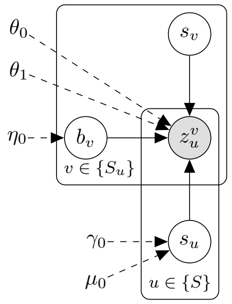
In the first model considered, PG_1 in [5] (Figure 2), variables are represented by circular nodes. The observed variables (in our case, only z_u^v) are shaded, while the rest are latent variables. The parameters of the prior distributions of the model are represented by Greek letters and dashed lines.



for $v \in \{S\}$:
  $\tau_v \sim \mathcal{G}(\alpha_0, \beta_0)$
  $b_v \sim \mathcal{N}(0, 1/\eta_0)$
for $u \in \{S\}$:
  $s_u \sim \mathcal{N}(\mu_0, 1/\gamma_0)$
for $u \in \{S\}$:
  for $v \in \{S_u\}$:
    $z_u^v \sim \mathcal{N}(s_u + b_v, 1/\tau_v)$

(a) (b)

Fig. 2: Model PG_1 [5]: (a) graph and (b) associated generative process. Each student, as peer-assessor, has individual precision and bias (τ_v and b_v, resp.). As a test taker, they get a real (unknown) grade, s_u. The grade assessed by peer v for student u depends on the actual performance of student u (s_u), and the precision τ_v and the bias b_v of student v as peer-assessor.



for $v \in \{S\}$:
  $b_v \sim \mathcal{N}(0, 1/\eta_0)$
for $u \in \{S\}$:
  $s_u \sim \mathcal{N}(\mu_0, 1/\gamma_0)$
for $u \in \{S\}$:
  for $v \in \{S_u\}$:
    $z_u^v \sim \mathcal{N}(s_u + b_v, 1/(\theta_1 \cdot s_v + \theta_0))$

(a) (b)

Fig. 3: Model PG_3 [5]: (a) graph and (b) associated generative process. Each student, as peer-assessor, has individual bias, b_v. As a test taker, they get a real (unknown) grade, s_u. The grade assessed by peer v for student u depends on the actual performance of student u (s_u), and the bias as peer-assessor b_v and the performance s_v as test-taker of student v. s_v and s_u are copies of the same variable for different students.

Each outer indexed box indicates that the variables it groups are repeated as many times as the index indicates. S is the total set of students, and S_u is the subset of students that peer-assess student u's work. This model assumes that each student shows their own level of precision and bias as peer-assessor, modeled by random variables τ_v and b_v, respectively. It also assumes that, if we knew the real grade s_u of student u, the grade assessed by peer v could be estimated given the bias b_v and some variability inversely proportional to the precision τ_v.

A second model, PG_3 in [5] (Figure 3), is considered. It explicitly models a key observation: peer-assessors are also students who took the test. This model assumes that the real grade of a student's test s_u is a reliable indicator of their precision as peer-assessors. Students also have, in their role as peer-assessors, their own bias term, b_v. Thus, if we knew the real grade s_u of student u, the grade assessed by peer v could be estimated given their bias b_v and some variability inversely proportional to the real grade s_v obtained by the peer-assessor v in the same test. Although shown duplicated in Figure 3, s_u and s_v represent the same variable: the real grade of a student (u or v) in the test.

In this work, we assume that the real grades s_u are the ones provided by the instructor, which might not always hold in a real class. The other model proposed in [5] (PG_2) is not considered. It requires several peer-assessed tests with the same class group to reuse acquired knowledge across tests, a scenario outside the scope of this study.

2) Bayesian approach:

There is limited information on the peer assessment framework. Only a small set of peer-assessed grades[4] is usually available. However, our models (Figs. 2 and 3) aim to learn students' characteristics to enable more robust estimations of the final grades, and data is required to do so. Instructors have long experience in grading their students and extensive prior knowledge about their behavior. Bayesian statistics allow us to use this prior knowledge as a reasonable starting assumption that counterbalances our current set of peer-assessed grades [6]. We codify these initial beliefs in the form of a priori probability distribution on the model parameters. At this time, the hypotheses about the shape of these parameters can be explicitly codified.

Bayesian inference consists of updating the initial beliefs with the information gathered from the observed data. One of its advantages is that it produces a probability distribution[5](posterior distribution) on the possible values of each parameter, which allows modeling the uncertainty around it.

Although PGMs allow efficient performing of inference, as the complexity of the model increases, exact inference becomes unfeasible. Approximate inference is the best alternative. It provides an answer to the query as a probability distribution that is *close* to the distribution that would actually answer the question. Approaches based on optimization (e.g., variational inference) or sampling (e.g., MCMC or Gibbs sampling) stand out. We use MCMC sampling from STAN [63], a *software* for statistical modeling and inference that does not require deriving a learning method for each specific model.

3) Selection criteria: A key element of our procedure is the technique that guides instructors in the choice of the next test to evaluate (line 4 in Alg. 1). It is reasonable to consider that an appropriate selection criterion would introduce information that enables the maximum reachable reduction of uncertainty. In our case, this is the real grade of the test that most reduces model uncertainty. Looking for a globally optimal solution for the supervision process (a whole ordered sequence of tests to supervise) requires the grades assigned by the instructor to be known in advance. In this work, the next test to supervise is selected at each step, and in this way, all our selection criteria can be considered as heuristics since we do not provide any guarantee that the composition of step-by-step decisions builds up to the global optimal solution.

Many techniques have been proposed in the context of active learning from crowds: different forms of entropy [56], [57], potential information gain [64], [65], uncertainty models based on the annotations or the learned model [56], [65], minimum variance [65], completely random strategies [56] or deterministic *round-robin* [57], [64]. Most of these techniques can be adapted to our problem. Partially inspired by such work, we have tested the following techniques, which use available data and the model in different ways:

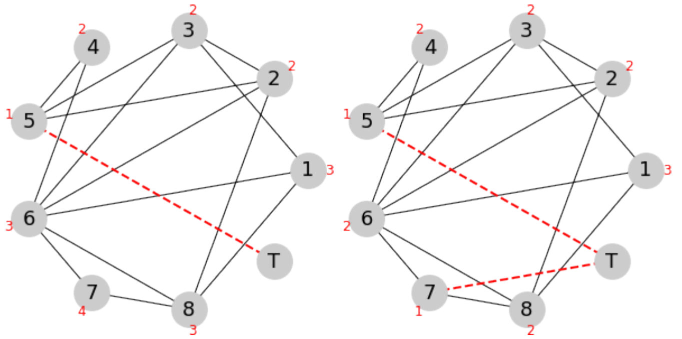


Fig. 4: Two steps of GpD criterion. The left figure displays student-instructor distances (red labels) as the instructor has only graded student #5's test. On the right, after the instructor grades the furthest student's test (#7), distances are updated.

a) Random selection (RND): The simplest alternative is entirely random selection of the next test to supervise. This does not use any source of information, but it could be close to the way an instructor acts in an unguided real scenario.

b) Graph distance (GpD): We can draw a graph that connects students using graded-by/grader-of relationships. Considering the general idea that information in a PGM flows through the network, the student that is furthest away from the instructor is impacted the least by the new information provided by the instructor. Thus, another straightforward baseline would be to select next the student who is furthest away from the instructor in the graph. See Figure 4 for a graphical example. There are multiple algorithms to find the distance between two nodes in a graph, and thus the student that is furthest *away* from the instructor. In this work, we measure distance in terms of the smaller number of edges that one needs to traverse to reach the target node from the source.

c) Grade variance (GdV): Each test is usually peer-assessed by multiple classmates. The variance among the peer-assessed grades can be understood as the degree of disagreement among peers. One could assign the tests to the instructor in decreasing-variance ordering, and thus the instructor will first examine those tests with the greatest divergence between the peer-assessed grades. These measurements are displayed in Figure 1 for our real dataset.

Following this approach comes with its own assumptions. Whereas a large variance can be fairly accepted as suspicious, a low variance is understood as correctness. This is hardly realistic. Peer assessments are subjective and students might agree to assign a specific grade to each other, or popular students can get their grades inflated. Anonymity over the whole peer-assessed task can alleviate these potential issues.

d) Posterior marginal variance (PMV): As we use Bayesian inference, when the instructor supervises a new test and provides the grade, the probability distributions on each

[4] Although one could ask each student to peer-assess all other students' tests to gather more data, that is not fair for students and it is definitively not in line with the formative objective of peer assessment.

[5] Unlike the frequentist approach, where data is used to estimate a specific value for the model parameters.

TABLE I: Parametrizations of PG_3 for synthetic data generation. Models 0 to 3 produce differences between real and peer-assessed grades in $[-2.5, 2.5]$. Models 4 to 7 reproduce the differences of Figure 1, in $[-0.75, 2.6]$. These differences may come from individual bias*, independent imprecision†, imprecision derived from lack of knowledge$^\circ$, or from all combined$^\triangleleft$. Other parameters are fixed to produce real grades around $\mu_0 = 7.5$ with a standard deviation of 1 ($\gamma_0 = 1$).

	Models							
Parameter	$0^\star$	$1^\dagger$	2°	$3^\triangleleft$	$4^\star$	$5^\dagger$	6°	$7^\triangleleft$
θ_0	12	0.8	0.01	0.04	15	1.1	0	0.1
θ_1	0.0	0.0	0.11	0.14	0.0	0.0	0.15	0.2
η_0	0.9	100	100	1.2	1.1	100	100	1.45
ρ_0	0.0	0.0	0.0	0.0	1.2	1.2	1.2	1.2

of the rest of the non-evaluated tests can be updated. The variance of the marginal distribution of each student's grade can be understood as the uncertainty associated with that test's grade. Note that in our models, the grades are modeled as normal distributions, and in this case entropy (a measure of uncertainty) and variance are directly coupled. With this technique, the next test suggested to the instructor will be the one with a higher posterior marginal variance in the model.

Unlike the previous technique, this relies on the model and this implies that the selection of the instructor in each step can modify the suggestion to the instructor in the next steps.

e) Maximum expected entropy reduction (MER): One could argue that the model is not used properly (or even not at all) by the previous techniques. From an *information gain* viewpoint [65], we aim to answer the question: "which is the test that the instructor needs to supervise next in order to reduce the total uncertainty in the model the most?".

We can calculate the entropy of the conditional posterior distribution of the yet-unknown grades, $\boldsymbol{S}$, given that we claim to know the value r of one of them, s_j : $h(\boldsymbol{S}\backslash\{s_j\})_{s_j=r}$. But, when we need to make the decision, we really do not know r. Thus, we resort to probabilistic estimates considering all possible values $r \in R$ for s_j (marginalization of s_j):

$$h_{s_j}(\boldsymbol{S}) = \int_{r\in R} p(s_j = r) \cdot h(\boldsymbol{S}\backslash\{s_j\})_{s_j=r} \qquad (1)$$

where R represents the set of all possible values of s_j, which in this study is a grade in the range from 0 to 10: $R = [0, 10]$.

Thus, we can calculate the expected entropy $h_{s_j}(\boldsymbol{S})$ for all the tests not yet supervised, $s_j \in \boldsymbol{S}$. Then, we would just need to suggest as the next test the student's test j that has associated a lower expected entropy, i.e., the one that promises the greatest reduction in the uncertainty in the model. In other words, it chooses the test that, once graded by the instructor, is expected to lead to a better estimate of the grades of the rest of the tests as yet unsupervised.

This is arguably the technique that best answers the question that leads the supervision procedure. However, it comes with a high price in terms of computational complexity[6].

To sum up, we consider 5 different selection techniques. First, a randomized approach is used to establish a baseline. We also include two approaches that use different types of data (GpD, which considers the relationships among students in the task, and GdV, which considers the peer-assessed grades), but they do not use the model for making the decision. Thus, we call them *data-based approaches*. Finally, two *model-based approaches* are included (PMV, which considers the uncertainty in the current estimates of the real grades, and MER, which estimates the maximum reduction of uncertainty if the instructor's grade of a certain test were known).

IV. Empirical study with synthetic data

Here we attempt to evaluate the viability of our proposal and test the suitability of the various models and techniques considered. To this end, two sets of experiments have been carried out. First, the results of a set of experiments performed with artificial data generated synthetically are presented in this section. Using synthetic data allows us to explore a much broader range of hypothetical (but reasonable) scenarios and evaluate the proposal and its different components in those conditions. Second, data from a controlled experiment in a real environment is used to evaluate the proposal in Section V.

A. Experimental design

Synthetic peer-assessed grades used in this set of experiments are generated to test different hypotheses. On the one hand, two possible scenarios for peer-assessed grades dispersion are explored: (i) differences between peer-assessed grades and *real* (unknown) grades are distributed normally in the interval $[-2.5, 2.5]$, assuming peers that over- and under-estimate grades in the same proportion, and (ii) differences normally distributed in the interval $[-0.75, 2.6]$, assuming peers that tend to overestimate grades in a larger proportion. The interval used in the latter case is calculated from the grades of the real dataset (Figure 1). On the other hand, we consider various hypotheses to explain the dispersion of peer-assessed grades. That is, we assume in distinct experiments that the differences between peer-assessed grades and *real* (unknown) grades are caused by (i) peer-assessors' individual bias, (ii) peer-assessors' independent lack of precision, (iii) peer-assessors' lack of precision derived from their lack of knowledge, or (iv) all three sources combined. In combination, up to 8 different hypothetical cases have been considered. We also use STAN for sampling the synthetic data from an enhanced PG_3 model, where an extra parameter ρ_0 represents a systemic bias, and thus b_v follows a normal distribution, $b_v \sim \mathcal{N}(\rho_0, 1/\eta_0)$. A complex data generation model allows us to accommodate all the assumptions and simulate realistic scenarios, as reality is always more complex than operating models. Parameter values have been selected by hand to simulate the 8 hypothetical cases (see Table I).

The two models presented in Section III-B1, PG_1 and PG_3, have been used to learn/model the simulated peer-assessment tasks, as well as three simplifications. We use a simplified model PG_1st where a single precision element is shared by all the students (in their role as peer-assessors), $\tau_v = \tau_{v'}, \forall v, v' \in$

[6] In our implementation, the grades are discretized as $\{[0, 5], (5, 6], (6, 7], (7, 8], (8, 9], (9, 10]\}$ during the calculation of Eq. 1 to alleviate the computational cost.

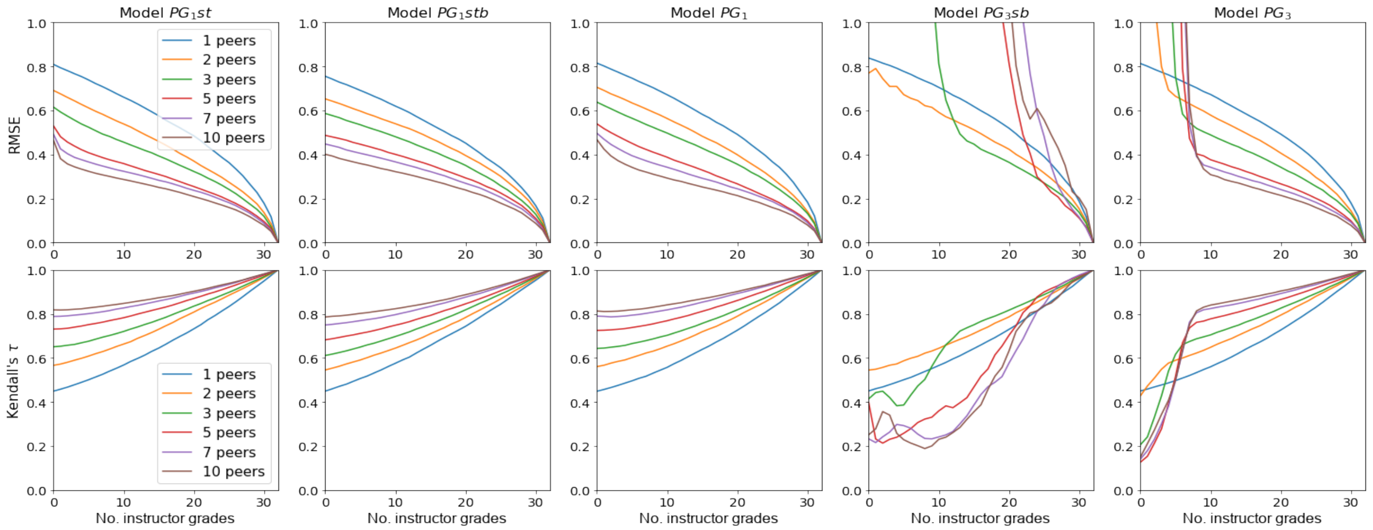


Fig. 5: Influence of the number of peer-assessments in terms of RMSE and *Kendall's* τ (by rows) when using different models (by columns). Performance evolution is shown as the instructor progresses in the supervision (from no test to all of them) where each line represents a different number of peer assessments per test, $G \in \{1, 2, 3, 5, 7, 10\}$. Results are averaged over all selection techniques and generative models and parameter $J = 32$ is fixed.

$\{S\}$. We restrict the expressiveness of PG_1 even further with another model, PG_1stb, where the bias element is also shared by all the students, $b_v = b_{v'}, \forall v, v' \in \{S\}$. Similarly, we use a model PG_3sb where the bias element is also shared by all the students, $b_v = b_{v'}, \forall v, v' \in \{S\}$. These last three simpler models will help us understand whether the data available in a peer-assessment task is enough to learn Piech et al. [5]'s models or whether more compact models are better suited for this task. The five techniques for selection of the next test to supervise (Section III-B3) have been considered: random selection (RND), graph distance (GpD), grade variance (GrV), posterior marginal variance (PMV), and maximum expected entropy reduction (MER).

Different class sizes have been simulated, $J \in \{8, 16, 24, 32, 40, 48, 56\}$, as well as different numbers of peer assessments per test, $G \in \{1, 2, 3, 5, 7, 10\}$. Only realistic configurations (where $G < J$) have been considered. The most informed selection technique, MER, has only been applied in experiments with $J \leq 24$ due to computational resource limitations. In total, 7240 experiments have been conducted in this exploratory design. Each experiment has been repeated 10 times to deal with the randomness of the synthetic data generation process. For the sake of clarity, all the plots show only averaged results, which allowing observation of the trend as the instructor advances in the revision.

To assess and compare the different techniques and models, we use the root mean square error (RMSE) and Kendall's rank correlation coefficient (Kendall's τ) [66]. RMSE is non-negative and measures the divergence between the ground truth values and their estimations, where a value of 0 identifies a perfect estimation. *Kendall's* τ is a rank correlation measure that, ignoring the actual values, compares two samples in that the values of both would be ordered similarly. It tends to 1 when the relative position of the students in the ordering of both samples (the real and estimated grades) is similar and tends to -1 when the order is rather the inverse. A value of 0 is expected when comparing two random samples. Whereas RMSE will help us identify how far our estimates differ from the ground truth, Kendall's τ tells us whether our estimations are actually giving a higher grade to whoever did better.

B. Results

The results presented below show different aspects of our empirical study to analyze the proposal's key elements.

In Figure 5 the evolution of the performance of the method, using the 5 models considered, is shown as the revision process carried out by the instructor progresses, paying special attention to the number of peer assessments per test ($G \in \{1, 2, 3, 5, 7, 10\}$) for a class of $J = 32$ students. In general, performance improves as the number of peer assessments per test increases. The difference is larger between tests with smaller numbers of peer assessments, and it seems to approach an upper bound as large values ($G = 7$ or 10) are used. The behavior with PG_1 and PG_1st is very similar and slightly better than that with PG_1stb. Performance with PG_3 is initially very unreliable but, as $20-25\%$ of the tests have been revised by the instructor, the behavior of the method using this model slightly outperforms PG_1-based approaches. A similar unstable start is observed for PG_3sb, but this is unable to achieve a competitive performance until very late in the revision process when many peer assessments are available.

Figure 6 shows the evolution of the performance of the method as the revision process goes on for different class sizes ($J \in \{8, 16, 24, 32, 40, 48, 56\}$). To be able to show the evolution for different class sizes, we plot it as a percentage of the total number of students in each experiment. It assumes a reasonable number of peer assessments per test ($G = 3$),

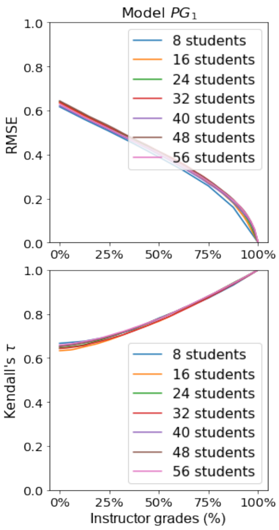


Fig. 6: Influence of the class size in terms of RMSE and *Kendall's* τ (by rows) when using PG_1 and PG_3 models (by columns). Performance evolution is shown as the instructor progresses in the supervision (from no test to all of them) where each line represents a different class size, $J \in \{8, 16, 24, 32, 40, 48, 56\}$. Results are averaged over all selection techniques and generative models. Parameter $G = 3$ is fixed.

and for the sake of simplicity only shows results for PG_1 and PG_3 models, as others follow the same trends. No difference is observed regarding class size. As before, PG_3 shows an initial unstable period which is quickly closed before $20-25\%$ of the tests have been revised by the instructor.

Regarding the influence of the selection criterion that guides instructors in the supervision, performance differences are very slight: MER, and PMV, to a lesser extent, appear to stand out in some scenarios (see a graphical description of these results in the supplementary material). Figure 7 shows the pairwise correlations between the order of tests suggested to instructors for revision when using different models. For the sake of simplicity, only results when following PMV and MER selection techniques are shown. GrV is deterministic (based exclusively on data) and shows full correlation for all the pairs of models, whereas RND is not expected to show any correlation due to complete randomization. GpD turns out to behave similarly to PMV. It can be observed that, when following PMV (or GpD), only spurious correlations are obtained between models. One could expect to observe a certain degree of correlation between models that are alike. Only MER shows this behavior: PG_3-based models have a strong correlation between them, also PG_1-based models, as well as a noticeable correlation between PG_1st and PG_3.

Finally, Figure 8 shows the behavior of the method when using the different models under specific generative assumptions. A standard class size of $J = 24$ students and $G = 3$ peer assessments per test are used. When data generative models only use bias as the source of differences between instructor (s) and peer-assessed grades (z), PG_1stb and PG_3sb show the worst performance as the models consider a single bias term common to all the students. PG_3 and PG_1 show a robust performance, but the best model in this scenario is PG_1st, which basically models bias and a single precision term common to all the students. When generative models use (lack of) precision as the single source of $s-z$ divergences, the differences between models are reduced, although the simplest models (PG_1stb and PG_3sb) show the best performance. Similar behaviors are observed when all students share the same precision term, or when each one has their own. When both bias and (lack of) precision are used as sources of $s - z$ differences, the more complete model PG_3 stands out. The initial unstable period of PG_3 (shorter) and PG_3sb (larger) is observed in all the cases, although the latter looks different when measured by RMSE or Kendall's τ.

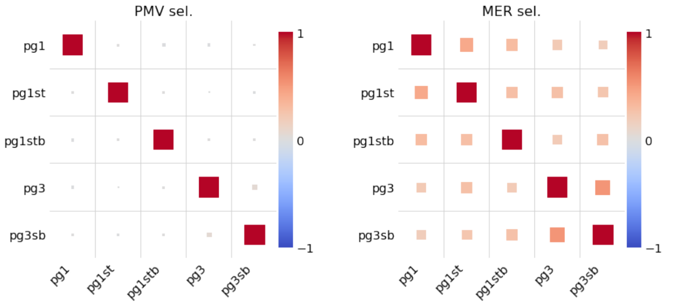


Fig. 7: Correlation between the order of tests suggested by PMV (left) and MER (right) criteria when modeling with the five different models across the whole set of experiments.

C. Discussion

Many insightful ideas can be extracted from the results concerning the behavior of the proposed method.

Primarily, it appears that class size has limited influence on the performance. According to Figure 6, the same proportion of tests revised by the instructor guarantees similar performance on the estimation of the remaining ones with different class sizes. This behavior, observed for all the models used, suggests that our proposal would apply to classes of any size, with a foreseeable workload for the instructor proportional to the class size.

As could be reasonably expected, the larger the number of peer assessments per test, the better the performance of the method. More data leads to better estimations in all the models used. This behavior is clearly observed in Figure 5. Similar behaviors have been previously reported, for example, by Conati *et al.* [37, p. 401], who claim that better performance can be obtained by "requiring students to display more of their thinking." This fact would motivate instructors to seek as many peer assessments as possible during the task. However, the workload of peer-assessing a test for a student, together with the time that it takes, is an obvious limitation on the number of peer assessments per test. It is the instructor's responsibility to estimate the number of tests that each student should peer assess without placing too much extra burden on them, possibly producing adverse effects such as loss of motivation, less exhaustive assessments, or even harm to the formative value of this evaluation methodology. In the real

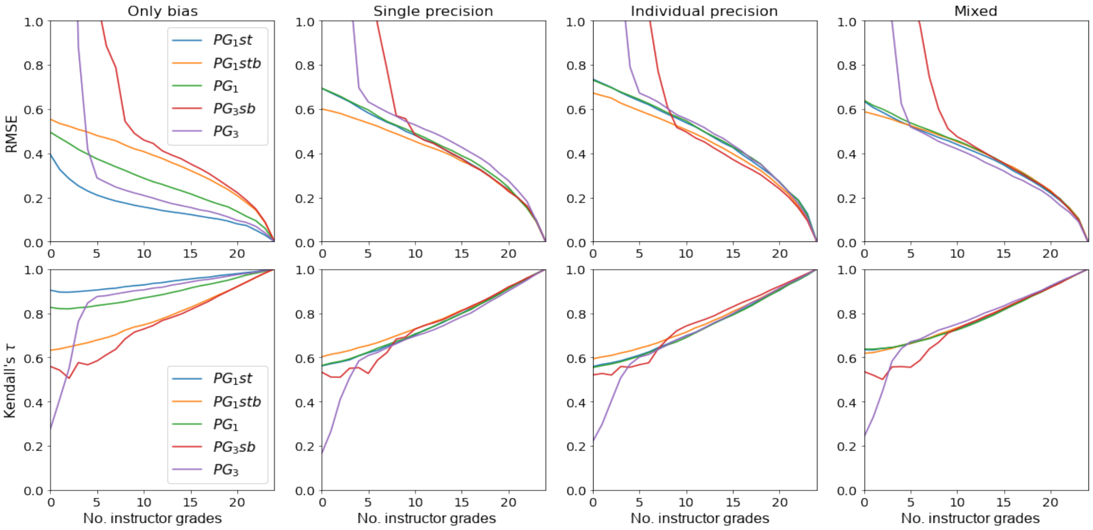


Fig. 8: Model performance when data is synthetically generated following different assumptions (Table I), by column. Measured in terms of RMSE and *Kendall's* τ (by rows), performance evolution is shown as the instructor progresses in the supervision (from no test to all of them). Results are averaged over all selection techniques and parameters $G = 3$ and $J = 24$ are fixed.

case study analyzed in this work (Sections III-A1 and V), each student performed and received 3 assessments.

In a data scarcity situation such as a peer-assessment task, simpler models can be learned more robustly. Thus, we compare PG_1 and PG_3 with 3 simplifications (in terms of the number of parameters), namely, PG_1st, PG_1stb and PG_3sb. PG_1-based models produce good estimations even with a small proportion of tests supervised by the instructor. The behavior of PG_1 and PG_1st is very similar across the whole empirical study, and the only observed difference is in experiments where the generative model clearly benefits PG_1st (see Fig. 8), which only models a bias element per student and a global precision term shared by all of them. The behavior of PG_1stb (it only learns a global bias and precision term) is slightly different: it is generally overtaken by the other two versions of PG_1 (clearly when the generative model relies on individual biases), and only shows slightly better performance when the generative model fixes a general bias term. With these results, it seems reasonable to limit the model by using a single precision term, but using a single bias term might be too restrictive. In other words, students tend to over/underestimate the grade of their peers differently, but they all are similarly consistent with their own biases. PG_3-based models have an initial unstable period during which they are outperformed by PG_1-based models. This period in the case of PG_3 lasts until the instructor has supervised not more than $20-25\%$ of tests. After this adjustment period, their performance can even surpass that of PG_1-based models. However, in the case of PG_3sb, this period is so long that it may hardly involve a workload reduction for the instructor. It is reasonable to expect instructors to supervise at least 25% of the tests to enhance validity, as previous evidence suggests [42]. In this sense, PG_3 is a competitive alternative to PG_1 and PG_1st. The failure of PG_3sb might be related to the apparently unrealistic assumption of using a single common bias term for all the students, also observed with PG_1stb.

Finally, the role of the selection technique appears limited in this empirical study. The differences between experiments that follow different selection techniques are negligible in most of the cases. The performance when following GrV, PMV or MER selection techniques is usually better than when using RND or GpD, which are indistinguishable between them. The correlation plots (similar to Fig. 7, but omitted for simplicity) show in both cases that no correlation exists between models (RND is fully randomized, and GpD hardly depends on an initial random selection). Exploring the *furthest* student first, in terms of graph distance, does not contribute to overcoming RND. GrD is deterministic once the peer-assessed grades are available, and thus, the correlation plot does not reveal anything about its performance. Surprisingly, PMV's correlation plot shows only spurious correlations between models. Only MER shows a partially predictable behavior, with a correlation plot that stresses the similarities of the path followed by PG_1-based and PG_3-based models. This and the slightly better performance of MER (in the supplementary material document) would suggest that this is arguably the most appropriate selection technique for our proposal.

V. Empirical Study with Real Data

Now we test our proposal with real data (Section III-A1).

A. Experimental design

For this analysis, peer-assessed grades for a test performed by $J = 16$ students are available. Each test received $G = 3$ grades. The instructor's grades are available too, $\{s_j\}_{i=1}^{J}$. This last piece of information is key as it allows us to carry the analysis out with the (assumed) ground truth information.

Similar to the previous experimental setting, five models (PG_1, PG_3, and their simplifications PG_1st, PG_1stb and PG_3sb) and five selection techniques (RND, GpD, GrV, PMV

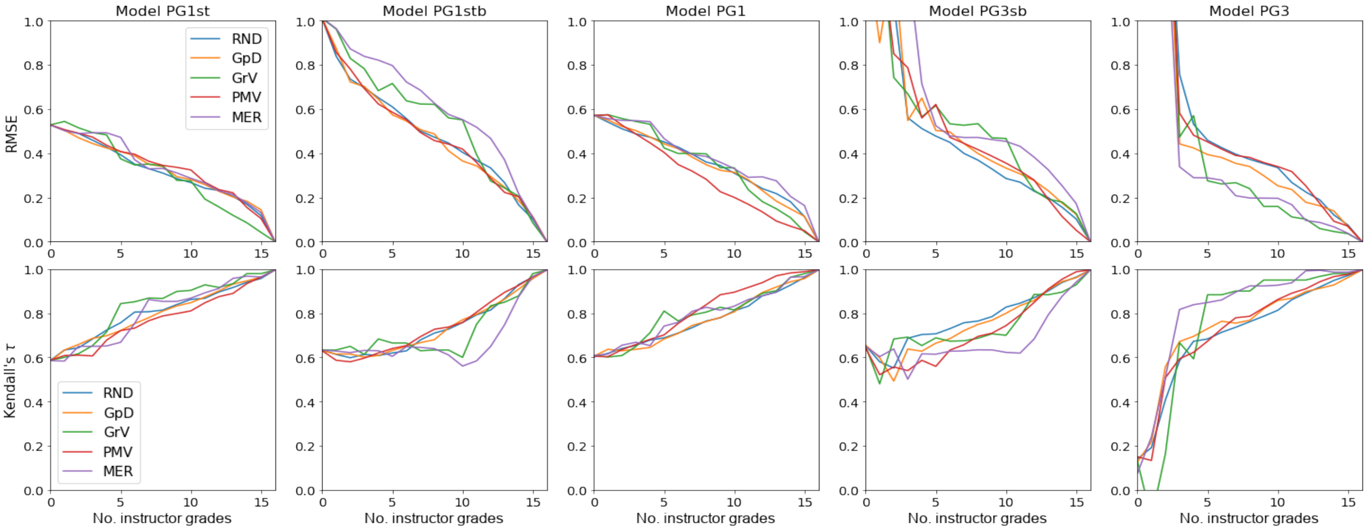


Fig. 9: Analysis of the influence of the selection technique per model on real data. Each subfigure shows the performance of five selection criteria (lines) in terms of RMSE and *Kendall's* τ (by rows) for five different learned models (by columns). The evolution of the metric is shown as the instructor evaluates more tests. Parameters $J = 16$ and $G = 3$ are fixed.

and MER) are tested. Each experiment is repeated 10 times to deal with the randomness of some selection techniques (in some cases, only the first selection). Average results are shown, in terms of RMSE and *Kendall's* τ, the metrics that we use to measure performance.

B. Results

Two different views of the analysis with real data are shown in Figures 9 and 10. Following a similar layout, two rows of subfigures show performance in terms of RMSE and Kendall's τ. In Figure 9, we group by model to compare the different selection techniques. In Figure 10, we group by selection technique to compare model performance.

The behavior of the procedure when using PG_1 and PG_1st is similar, although the best selection criteria with PG_1 is PMV, and GrV with PG_1st. The model leading to the worst behavior of the procedure is PG_1stb. With it, the performance limitations of MER and GrV are noticeable (mainly in terms of RMSE). A similar performance drop is observed between PG_3 and PG_3sb, with MER and GrV especially impaired when using single bias modeling assumption. Note that for PG_3sb, sometimes RND is the best selection criterion. Experiments with PG_3 (and with PG_3sb in terms of RMSE) show initially very poor performance, but with competitive behavior attained quickly once the instructor has revised a few tests. In fact, when using the MER (or GrV) criteria, the best performance is reached earliest (with the fewest revised tests).

When grouped by selection technique, the procedure behaves similarly if the RND or GpD criteria are applied. The behavior when following PMV is also quite similar, although some differences are observed (e.g., PG_1 is benefited). Under these criteria, modeling with PG_1 or PG_1st would be the best decision, although PG_3 is also competitive after several evaluations of the instructor. PG_3sb and PG_1stb are the worst models for all the selection techniques, almost always, in terms of both RMSE and Kendall's τ. The behavior varies the most across models when following GrV or MER selection criteria. Under these criteria, PG_3 stands out as the model that leads to the best results, followed by PG_1st and PG_1.

C. Discussion

These experiments with real data show larger differences when all the models and selection techniques are compared.

The behavior of the method when using the different models diverges, as can be seen in Figure 9. PG_1 and PG_1st show similar behaviors, although the leading selection technique in each case is different. In fact, PG_1st outperforms PG_1 when following 4 out of 5 selection techniques (see Fig. 10), including GrV and MER. This strengthens the hypothesis behind PG_1st in a data-scarce scenario such as ours: all the students have their own bias and diverge from it randomly in the same way. Those models that consider a single/common bias term for all the students, PG_1stb and PG_3sb, show the worst performance, mainly in terms of RMSE. In line with experiments in synthetic data, at least in this real case study, the single-bias assumption appears to be too unrealistic. Both PG_3-based models again show the initial adjustment period before the instructor has supervised up to $20 - 25\%$ of the tests. After this adjustment period, PG_3 stands out as the best model when following MER and GrV selection techniques, with a performance improvement compared to other models of about 35%. These results, consistent with others from the previous Section IV, support the modeling decision that connects students' roles of test-taker and peer-assessor, assuming that whoever does it well when being evaluated, also does well as a peer assessor.

Unlike the empirical study with synthetic data (Sec. IV), in this case, there are considerable differences between the

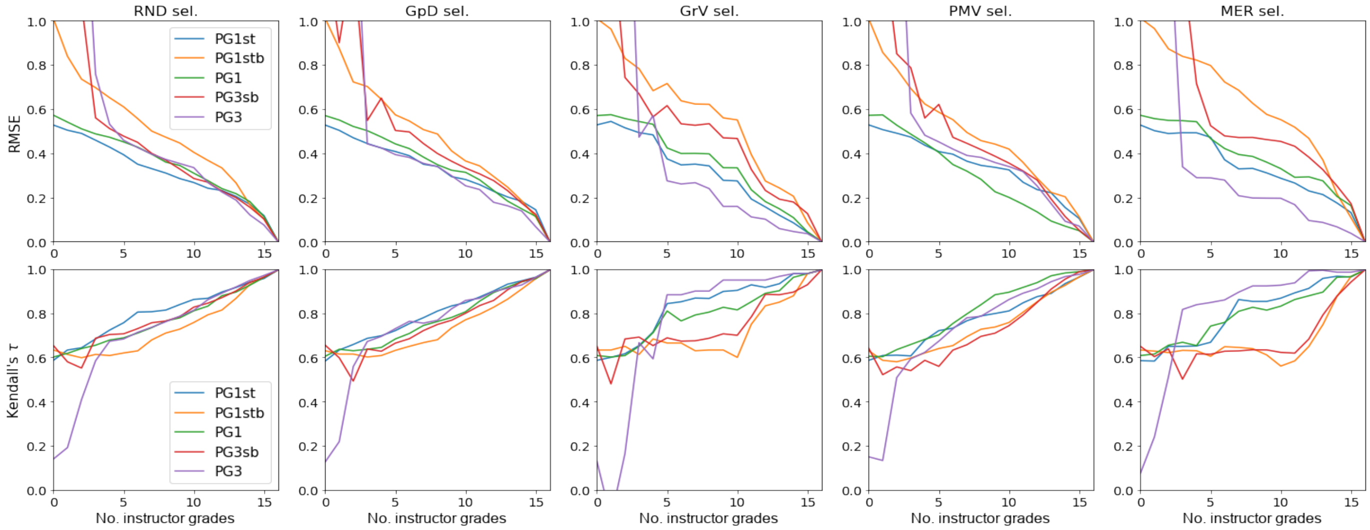


Fig. 10: Analysis of the influence of the model per selection criteria on real data. Each subfigure shows the performance of five models (lines) in terms of RMSE and *Kendall's* τ (by rows) for five different selection criteria (by columns). The evolution of the metric is shown as the instructor evaluates more tests. Parameters $J = 16$ and $G = 3$ are fixed.

selection techniques. Procedure behavior is very similarly when following the RND and GpD selection criteria (and even with PMV) across all five models (see Fig. 10). This common behavior was already observed in the previous Section IV, and it could be arguably attributed to the near-random behavior of these selection techniques (see Fig. 7). The GrV and MER selection techniques induce large differences between models. In both cases, PG_3 stands out (after its adjustment period), followed by PG_1st and PG_1 (in that order). Finally, PG_3sb and PG_1stb never show competitive performance. One might think that the differences between these results and those of Section IV are due to unrealistic synthetic data or similar issues. However, the results from this section should be analyzed with caution as they come from a single task.

All in all, the most robust configuration of our proposal would be to use the PG_3 model following the entropy-based selection technique, MER. In this case, it must be stressed that to overcome the adjustment period of this model, the instructor should review at least $20-25\%$ of the tests. It is worth noting that de Alfaro and Shavlovsky [42] found while analyzing the use of an online platform for peer assessment that instructors spontaneously decided to supervise roughly 30% of the tests. Alternatively, the use of the PG_1st model following GrV also shows a competitive performance, with even better results in the initial steps.

VI. Threats to validity

This study assumes that both instructors and students can potentially benefit from the proposal: a workload reduction for instructors, and increased validity of aggregated grades for both. We used anonymous questionnaires for checking whether these assumptions are realistic and to gather opinions of both groups (the form and raw results are available on the associated webpage). We gathered 45 and 22 answers from instructors and students, respectively. According to instructors' perceptions, the expected validity of peer-assessed grades increases by 0.82 points (on a scale from 1 to 5) using our method. This increase reaches 1.17 points when asking only those instructors who have previously used peer assessment. Among students, the expected validity increase is 0.68 and 0.46 points, respectively. Over 60% of instructors would use peer assessment for grading but only 9% would rely completely on peer-assessed grades. 51% of instructors prefer supervising the task when using peer assessment for grading. All in all, 76% of instructors appear open to using the method (chiefly, after learning how it works). 64% of students appear open to peer assessment and 59% accept the use of peer assessment for grading (mainly with instructor supervision). The main concerns are a lack of trust in classmates ("sometimes it is not about the peers' malicious intention, it is just a matter of lack of ability or commitment") and the feeling of carrying out a task (assessment) that should be the instructor's responsibility. In contrast, 73% are open to the use of our method if it is explainable and controlled by the instructor ("aggregation should not be completely automatized"). These opinions gathered appear to reveal an overall positive opinion toward the benefits of using a method such as ours.

In this study, the instructor's grade is assumed to be the real grade, which is not necessarily true. This limitation could be addressed by enlarging the models with separate random variables for the instructor and real grades. All our models use a single numerical grade per test. We could check how our methodology combines with other models that allow lower-granularity grades (e.g., grade per rubric item) and categorical or ordinal grades [41], [45]. Our Bayesian approach involves the selection of priors. The hyperparameters are set to uninformative values, or they could be set to summary statistics of the available data. Previous works claim that these models could be not so sensitive to the specific values of the hyperparameters [37]. A specific study is necessary to claim so for our method. The selection technique MER, which appears to outperform the rest, cannot be used extensively due to its large computational cost. Alternative approximate inference techniques such as variational inference should be explored. Moreover, the selection of the two evaluation metrics (RMSE and Kendall's τ) arguably fulfills our initial intention of displaying different views (quantitative and qualitative) of the results. Experimental results are different regarding both metrics but consistent, which supports our selection.

The empirical study with real data analyses a middle-sized class. Although our synthetic-data experimental results suggest that our method is not class-size dependent, other sizes should be tested. It was performed in a university class. A larger study is required to test whether these results generalize to other classes at the same level and at other educational levels. The peer-assessed test was an extended-response exam consisting of three questions. A specific study should investigate whether these results generalize to a broader range of exam types. The rubric employed needs to be adjusted for the students, as peers and instructors may interpret them differently [22]. After the rubric is used, it should be revised using, for example, Kulkarni et al.'s methodology [22]: (i) testing the validity of the peer-assessed grades at a rubric-item granularity, and (ii) improving the wording of the low-validity items. The test was carried out in the context of anonymity. A larger study could test our methodology when this condition does not hold and preference relationships between students bias the peer-assessed grades. Presumably, our method will require models that account for social interactions [47].

Regarding the study with synthetic data, the performance of our technique could have benefited from the use of models that match the data generation process. Empirical results with real data provide initial evidence against the occurrence of this threat. However, it is undeniable that the configuration of the data generation process follows the real data characteristics. A larger study should consider other artificial scenarios, for example, using popularity models to simulate specific students being treated differently in peer assessment tasks.

VII. Conclusions

This paper presents a method to assist the instructor in the supervision of a peer-assessment task. It combines a Bayesian probabilistic graphical model to model the peer-assessment task, and a selection technique to suggest to the instructor which task they should evaluate next. The estimation of the uncertainty associated with the model enables a selection criterion based on minimum expected uncertainty, and the early cessation of the supervision in line with the instructor's own tolerable uncertainty criteria. It is well-known that peer assessment has a positive impact on student training and our procedure might help spread its adoption, given that it leads to a reduction in instructors' workload and provides confidence in the form of mathematically well-founded grade estimates. Finally, any difficult decision is always made by the instructor.

The proposed methodology has been validated with synthetic and real data. Graphical models carrying different modeling assumptions and different selection techniques (from simple to elaborated) have been studied. Results appear to indicate that students have their own bias but indistinguishable precision, and that whoever does it well as a test taker, does well as a peer assessor. The selection technique that stands out estimates the uncertainty as if the real grade of each test was known and suggests the one with the largest estimated entropy reduction. However, a computationally-cheap and competitive alternative is to follow the order given by the variance of the peer-assessed grades. Two surveys carried out among instructors and students, respectively, show promising potential acceptance of the use of our methodology.

Acknowledgments

This work was supported in part by the UNED through Grant GID2016-6, and in part by the Community of Madrid and EU's Structural Funds through Grant S2018/NMT-4331.

References

[1] N. Falchikov and J. Goldfinch, "Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks," *Rev. Educ. Res.*, vol. 70, no. 3, pp. 287–322, 2000.

[2] S. San Martín Gutiérrez, N. Jiménez Torres, and E. J. Sánchez-Beato, "La evaluación del alumnado universitario en el Espacio Europeo de Educación Superior," *Aula Abierta*, vol. 44, no. 1, pp. 7–14, 2016.

[3] C. Brindley and S. Scoffield, "Peer assessment in undergraduate programmes," *Teach. High Educ.*, vol. 3, no. 1, pp. 79–90, 1998.

[4] Y. Bachrach, T. Minka, J. Guiver, and T. Graepel, "How to grade a test without knowing the answers- a Bayesian graphical model for adaptive crowdsourcing and aptitude testing," in *Proc. of 29th ICML*, 2012.

[5] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller, "Tuned models of peer assessment in MOOCs," in *Proc. of 6th EDM*, 2013.

[6] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[7] B. Settles, "Active learning literature survey," University of Wisconsin Madison, Tech. Rep. no. 1648, 2009.

[8] K. J. Topping, "Peer assessment between students in colleges and universities," *Rev. Educ. Res.*, vol. 68, no. 3, pp. 249–276, 1998.

[9] W. Harlen and M. James, "Creating a positive impact of assessment on learning," in *Proc. of An. Mtg. AERA*, 1996, pp. 2–12.

[10] I. Ruz Herrera, "Evaluación para el aprendizaje," *Rev. Educ. Las Américas*, vol. 6, pp. 13–28, 2018.

[11] J. Gil Flores and M. T. Padilla Carmona, "La participación del alumnado universitario en la evaluación del aprendizaje," *Educación XX1*, 2009.

[12] P. M. Sadler and E. Good, "The impact of self- and peer-grading," *Educ. Assess.*, vol. 11, no. 1, pp. 1–31, 2006.

[13] D. Reinholz, "The assessment cycle: A model for learning through peer assessment," *Assess. Eval. High. Educ.*, vol. 41, no. 2, pp. 301–315, 2016.

[14] S. T. Basurto-Mendoza, J. A. Moreira-Cedeño, A. N. Velásquez-Espinales, and M. Rodríguez-Gámez, "Autoevaluación, coevaluación y heteroevaluación como enfoque innovador en la práctica pedagógica y su efecto en el proceso de enseñanza-aprendizaje," *Polo del Conocimiento*, vol. 6, pp. 828–845, 2021.

[15] L. Bardach, R. M. Klassen, T. L. Durksen, J. V. Rushby, K. C. Bostwick, and L. Sheridan, "The power of feedback and reflection: Testing an online scenario-based learning intervention for student teachers," *Comput. Educ.*, vol. 169, p. 104194, 2021.

[16] A. Membrive Ruiz, M. Largo Sierra, C. Cáceres, M. I. Vizquerra Fletcher, A. Engel Rocamora, and M. Solari, "La reflexión como estrategia de personalización del aprendizaje escolar," in *Proc. of 1st I. Conf. Recerca en Educació*, 2020, pp. 874–882.

[17] K. Misiejuk, B. Wasson, and K. Egelandsdal, "Using learning analytics to understand student perceptions of peer feedback," *Comput. Hum. Behav.*, vol. 117, p. 106658, 2021.

[18] T. Hovardas, O. E. Tsivitanidou, and Z. C. Zacharia, "Peer versus expert feedback: An investigation of the quality of peer feedback among secondary school students," *Comput. Educ.*, vol. 71, pp. 133–152, 2014.

[19] C. Canabal and L. Margalef, "La retroalimentación: La clave para una evaluación orientada al aprendizaje," *Profesorado*, vol. 21, pp. 149–170, 2017.

[20] M. A. Gómez Ruíz and V. Quesada Serra, "Coevaluación o evaluación compartida en el contexto universitario: La percepción del alumnado de primer curso," *Rev. Iberoam. Eval. Educ.*, vol. 10, no. 2, 2017.

[21] H. G. Andrade, "Teaching with rubrics: The good, the bad, and the ugly," *Coll. Teach.*, vol. 53, no. 1, pp. 27–31, 2005.

[22] C. Kulkarni, K. Wei, H. Le, D. Chia, K. Papadopoulos, J. Cheng, D. Koller, and S. Klemmer, "Peer and self assessment in massive online classes," *ACM Trans. Comput.-Hum. Interact.*, vol. 20, no. 6, 2013.

[23] V. M. López Pastor and A. Pérez Pueyo, *Evaluación formativa y compartida en educación: Experiencias de éxito en todas las etapas educativas*. Universidad de León, 2017.

[24] A. Torregrosa, L. Albarracín, and J. Deulofeu, "Orientación y coevaluación: Dos aspectos clave para la evolución del proceso de resolución de problemas," *Bolema*, vol. 35, pp. 89–111, 2021.

[25] N. Sanmartí, *Avaluar i aprendre: Un únic procés*. Octaedro, 2019.

[26] E. Barrientos Hernán, V. M. López Pastor, and D. Pérez-Brunicardi, "¿Por qué hago evaluación formativa y compartida y/o evaluación para el aprendizaje en EF? La influencia de la formación inicial y permanente del profesorado," *Retos*, vol. 36, pp. 37–43, 2019.

[27] I. Álvarez Valdivia, "La coevaluación como alternativa para mejorar la calidad del aprendizaje de los estudiantes universitarios: Valoración de una experiencia," *Rev. Interuniv. Formac. Prof.*, vol. 22, pp. 127–140, 2008.

[28] A. Gessa Perera, "La coevaluación como metodología complementaria de la evaluación del aprendizaje. Análisis y reflexión en las aulas universitarias," *Rev. Educ.*, vol. 354, pp. 749–764, 2011.

[29] H. Luo, A. C. Robinson, and J. Y. Park, "Peer grading in a MOOC: Reliability, validity, and perceived effects," *J. Async. Learn. Netw.*, vol. 18, no. 2, pp. 1–14, 2014.

[30] H.-H. Chang, C. Wang, and S. Zhang, "Statistical applications in educational measurement," *An. Rev. Stat. Appl.*, vol. 8, no. 1, pp. 439–461, 2021.

[31] B. Van De Sande, "Properties of the Bayesian knowledge tracing model," *J. Educ. Data Min.*, vol. 5, no. 2, pp. 1–10, 2013.

[32] Y. In'nami and R. Koizumi, *Structural equation modeling in educational research: A primer*. Brill, 2013, pp. 23–51.

[33] N. Denson and M. Ing, "Latent class analysis in higher education: An illustrative example of pluralistic orientation," *Res. High. Educ.*, vol. 55, no. 5, pp. 508–526, 2014.

[34] R. J. Harvey and A. L. Hammer, "Item response theory," *Couns. Psychol.*, vol. 27, no. 3, pp. 353–383, 1999.

[35] R. Almond, R. Mislevy, L. Steinberg, D. Yan, and D. Williamson, *Bayesian networks in educational assessment*. Springer, 2015.

[36] R. J. Mislevy, R. G. Almond, D. Yan, and L. S. Steinberg, "Bayes nets in educational assessment: Where the numbers come from," in *Proc. of 15th UAI*, 1999, pp. 437–446.

[37] C. Conati, A. Gertner, and K. VanLehn, "Using Bayesian networks to manage uncertainty in student modeling," *User Model. User-Adapt. Interact.*, vol. 12, no. 4, pp. 371–417, 2002.

[38] M. J. Culbertson, "Bayesian networks in educational assessment: The state of the field," *Appl. Psychol. Meas.*, vol. 40, no. 1, pp. 3–21, 2016.

[39] I. Uglanova, "Model criticism of Bayesian networks in educational assessment: A systematic review," *Pract. Assess. Res. Eval.*, vol. 26, 2021.

[40] R. Levy, "Advances in Bayesian modeling in educational research," *Educ. Psychol.*, vol. 51, no. 3-4, pp. 368–380, 2016.

[41] K. Raman and T. Joachims, "Methods for ordinal peer grading," in *Proc. of 20th ACM KDD*, 2014, pp. 1037–1046.

[42] L. de Alfaro and M. Shavlovsky, "Dynamics of peer grading: An empirical study," in *Proc. of 9th EDM*, 2016, pp. 62–69.

[43] B. Shalem, Y. Bachrach, J. Guiver, and C. M. Bishop, "Students, teachers, exams and MOOCs: Predicting and optimizing attainment in web-based education using a probabilistic graphical model," in *Proc. of ECML/PKDD 2014*, 2014, pp. 82–97.

[44] T. Wang, X. Jing, Q. Li, J. Gao, and J. Tang, "Improving peer assessment accuracy by incorporating relative peer grades," in *Proc. of 12th EDM*, 2019, p. 450–455.

[45] F. Mi and D.-Y. Yeung, "Probabilistic graphical models for boosting cardinal and ordinal peer grading in MOOCs," in *Proc. of 29th AAAI*, 2015, p. 454–460.

[46] H. Zarkoob, G. d'Eon, L. Podina, and K. Leyton-Brown, "Better peer grading through Bayesian inference," 2022, arXiv 2209.01242.

[47] H. P. Chan and I. King, "Leveraging social connections to improve peer assessment in MOOCs," in *Proc. of 26th WWW*, 2017, p. 341–349.

[48] J. Xu, Q. Li, J. Liu, P. Lv, and G. Yu, "Leveraging cognitive diagnosis to improve peer assessment in MOOCs," *IEEE Access*, vol. 9, pp. 50 466–50 484, 2021.

[49] R. Reichenberg, "Dynamic Bayesian networks in educational measurement: Reviewing and advancing the state of the field," *Appl. Meas. Educ.*, vol. 31, no. 4, pp. 335–350, 2018.

[50] J. Howe, "The rise of crowdsourcing," *Wired Magazine*, vol. 15, no. 6, pp. 1–4, 2006.

[51] J. Zhang, X. Wu, and V. S. Sheng, "Learning from crowdsourced labeled data: A survey," *Artif. Intell. Rev.*, vol. 46, no. 4, pp. 543–576, 2016.

[52] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *J. R. Stat. Soc. Ser. C-Appl. Stat.*, vol. 28, no. 1, pp. 20–28, 1979.

[53] V. C. Raykar, S. Yu, L. H. Zhao, G. Hermosillo Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *J. Mach. Learn. Res.*, vol. 11, pp. 1297–1322, 2010.

[54] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. R. Movellan, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Proc. of 22sd NeurIPS*, 2009, pp. 2035–2043.

[55] P. Welinder, S. Branson, S. Belongie, and P. Perona, "The multidimensional wisdom of crowds," in *Proc. of 23 NeurIPS*, 2010, pp. 2424–2432.

[56] V. S. Sheng, F. J. Provost, and P. G. Ipeirotis, "Get another label? Improving data quality and data mining using multiple, noisy labelers," in *Proc. of 14th ACM KDD*, 2008, pp. 614–622.

[57] M. Venanzi, O. Parson, A. Rogers, and N. Jennings, "The active crowd toolkit: An open-source tool for benchmarking active learning algorithms for crowdsourcing research," in *Proc. of 3rd HCOMP*, 2015.

[58] P. Welinder and P. Perona, "Online crowdsourcing: Rating annotators and obtaining cost-effective labels," in *Proc of 23rd CVPR Workshops*, 2010, pp. 25–32.

[59] N. Q. V. Hung, D. C. Thang, M. Weidlich, and K. Aberer, "Minimizing efforts in validating crowd answers," in *Proc. of ACM SIGMOD/PODS*, 2015, pp. 999–1014.

[60] Y. Yan, R. Rosales, G. Fung, and J. G. Dy, "Active learning from crowds," in *Proc. of 28th ICML*, 2011, pp. 1161–1168.

[61] J. Vomlel, "Bayesian networks in educational testing," *Int. J. Uncertain. Fuzz. Knowl.-Based Syst.*, vol. 12, no. supp01, pp. 83–100, 2004.

[62] E. Millán, M. Trella, J.-L. Pérez-de-la Cruz, and R. Conejo, "Using Bayesian networks in computerized adaptive tests," in *Computers and Education in the 21st Century*. Springer, 2000, pp. 217–228.

[63] S. D. Team, "STAN modeling language users guide and reference manual, 2.26," Software, 2021. [Online]. Available: https://mc-stan.org

[64] A. R. Khan and H. Garcia-Molina, "CrowdDQS: Dynamic question selection in crowdsourcing systems," in *Proc. of ACM SIGMOD/PODS*, 2017, pp. 1447–1462.

[65] Y. Lewenberg, Y. Bachrach, U. Paquet, and J. S. Rosenschein, "Knowing what to ask: A Bayesian active learning approach to the surveying problem," in *Proc. of 31st AAAI*, 2017, pp. 1396–1402.

[66] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1-2, pp. 81–93, 1938.

Jerónimo Hernández-González is a tenure-eligible lecturer in the department of Mathematics and Computer Science at the University of Barcelona, Spain. He received his PhD from the University of the Basque Country, Spain, in 2015. His major research interests include learning and inference with probabilistic graphical models and their application to biomedicine and education.

Pedro Javier Herrera is currently an Associate Professor with the Software Engineering and Computer Systems Department, National University of Distance Education (UNED), Spain. He received his PhD in computer science from the Complutense University of Madrid, Spain, in 2010. His research activities include computer vision, pattern recognition, artificial intelligence, and robotics with interests in continuous training and teaching innovation.