Using Sampling to Estimate and Improve Performance of Automated Scoring Systems with Guarantees

Yaman Kumar Singla[1,2,3*], **Sriram Krishna**[1*], **Rajiv Ratn Shah**[1], **Changyou Chen**[3]

[1]IIIT-Delhi, [2]Adobe Media Data Science Research, [3]State University of New York at Buffalo
ykumar@adobe.com, sriramsk1999@gmail.com, rajivratn@iiitd.ac.in, changyou@buffalo.edu

Abstract

Automated Scoring (AS), the natural language processing task of scoring essays and speeches in an educational testing setting, is growing in popularity and being deployed across contexts from government examinations to companies providing language proficiency services. However, existing systems either forgo human raters entirely, thus harming the reliability of the test, or score every response by both human and machine thereby increasing costs. We target the spectrum of possible solutions in between, making use of both humans and machines to provide a higher quality test while keeping costs reasonable to democratize access to AS. In this work, we propose a combination of the existing paradigms, sampling responses to be scored by humans intelligently. We propose reward sampling and observe significant gains in accuracy (19.80% increase on average) and quadratic weighted kappa (QWK) (25.60% on average) with a relatively small human budget (30% samples) using our proposed sampling. The accuracy increase observed using standard random and importance sampling baselines are 8.6% and 12.2% respectively. Furthermore, we demonstrate the system's model agnostic nature by measuring its performance on a variety of models currently deployed in an AS setting as well as pseudo models. Finally, we propose an algorithm to estimate the accuracy/QWK with statistical guarantees. Our code is available at https://git.io/J1IOy.

1 Introduction

Automated Scoring (AS), the task of assigning scores to unstructured responses to open-ended questions, is an NLP application typically deployed in an educational setting. Historically, its origins have been traced to the work of Ellis Page (Page 1967), who first argued for the possibility of scoring essays by computer. The factors behind the rise of Automated Scoring systems and its subtasks, Automated Essay Scoring (AES) and Automated Speech Scoring (ASS) are numerous, including but not limited to, the costs involved in providing and scoring a test, and ensuring that all test takers are scored on a uniform set of rubrics applied across all students, standardizing the scoring for these unstructured responses. The promise of lower costs and uniform scoring rubrics among other factors, has fueled the popularity of Automated Scoring systems, and various ML and DL systems are being increasingly deployed in AS contexts (Kumar et al. 2019; Liu, Xu, and Zhu 2019; Singla et al. 2021a). AS systems are behind some of the world's most popular language tests, such as ETS' Test of English as a Foreign Language (TOEFL) (Zechner et al. 2009), Duolingo's English Test (DET) (LaFlair and Settles 2019), among others. Various governmental institutions and businesses have also instituted automated systems to augment the scoring process, such as the state schools of Utah (Incorporated 2017) and Ohio (O'Donnell 2018), and a majority of BPOs. It is estimated that automatic scoring has a large market size of more than USD 110 billion, with a US market size alone of USD 17.1 billion (TechNavio 2020; Service 2020; Strauss 2020; Le 2020).

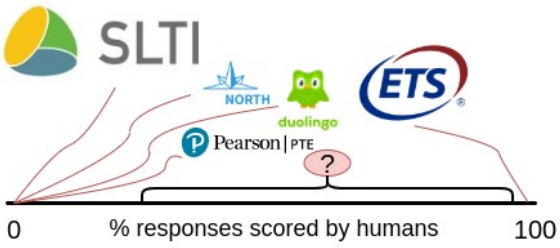


Figure 1: Existing Automated Scoring systems[1] either do not involve humans at all in their scoring (Duolingo, Second Language Testing Inc (SLTI)), or utilize human raters for every single response (Educational Testing Services (ETS)). Crucially, there are no solutions that target the gulf in between, where humans are involved in scoring only some percentage of the responses.

However, this popularity has not been without backlash, with criticism focusing on different aspects, such as "the overreliance on surface features of responses, the insensitivity to the content of responses and to creativity, and the vulnerability to new types of cheating and test-taking strate-

*These authors contributed equally.

[1]TOEFL by ETS, Pearson PTE, SLTI, Duolingo English Test, and TrueNorth by Emmersion are registered brand names and are shown here for illustration purposes only. The authors claim no rights over their logos or brand names. In this work, we mainly refer to the automatically scored speaking and writing proficiency measurement tests of these companies.

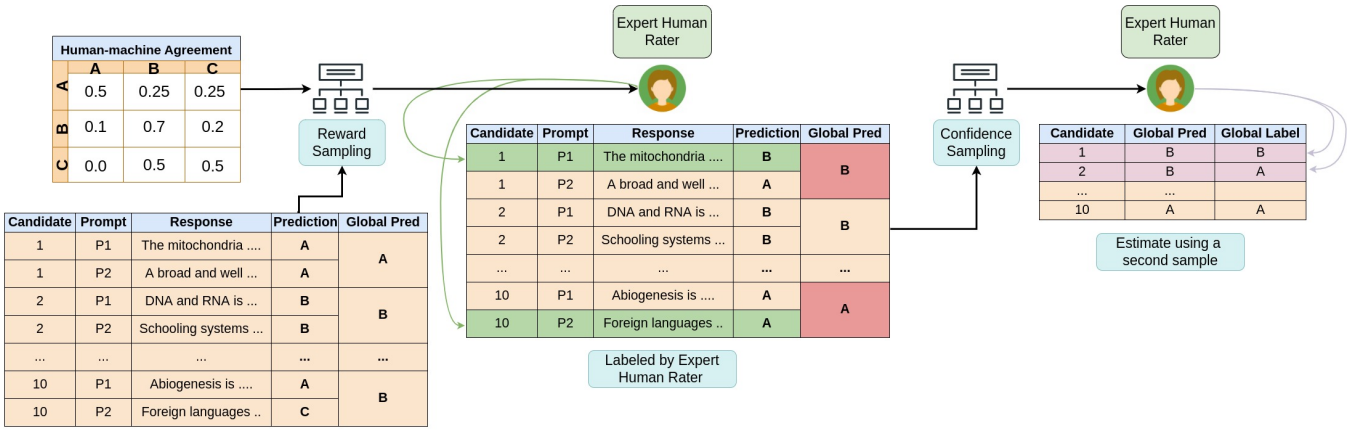


Figure 2: From a dataset, records are sampled and assigned to expert human raters for double scoring based on a *human-machine agreement matrix*. A second sample is then drawn to check predictions and metrics are estimated with statistical guarantees.

gies." (Yang et al. 2002). Others have given harsher criticisms, such as (Perelman et al. 2014), who shows that it is possible to *game* the system and achieve near perfect scores on ETS and Vantage Technologies' AES systems with gibberish prose. This has led to the revoking of NAPLAN AES in Australia (ACARA 2018).

Nonetheless, the ability of AS systems to instantly provide scores, reduce costs, and make language proficiency tests more widely available to all, makes them an important research area and subsequently there is considerable interest in improving them across multiple dimensions, from leveraging advancements in NLP to achieve state-of-the-art performance (Liu, Xu, and Zhu 2019) to improving their robustness (Kumar et al. 2020; Parekh et al. 2020; Singla et al. 2021b). In this work, we tackle another facet of Automatic Scoring systems, that of improving performance by bringing humans into the loop.

Typically in an AS task, a test taker's responses are scored on prompts of varying difficulty levels. Each prompt has its own difficulty level, and based on the prompts' difficulty and the quality of the candidate's answers to these prompts, a score is assigned to the candidate. The Central European Framework of Reference for Languages (CEFR) is an international standard for measuring language proficiency and assigns scores on a six-point scale from **A1** (beginner) to **C2** (proficient), each score with their own rubrics for evaluation (Broeder and Martyniuk 2008). Each prompt and response is assigned a score on this scale and a **global** score is computed aggregating these individual scores.

Existing AS systems are typically of two varieties (Fig 1):

Double Scoring: Examinations such as ETS' TOEFL score every response by one human and an AES system as the second rater. A second human rater resolves any disagreements between the two (Yang et al. 2002). This effectively means that atleast one human rater is required for every test, driving up costs, as evidenced by the TOEFL's high price of ∼230 USD (ETS 2021).

Machine-only Scoring: On the other end of the scale are tests like the Duolingo English Test (DET) which are scored by machines alone, without any human intervention, keeping costs low but decreasing the reliability of the test. This is one of the main reasons, the DET costs USD 49, less than one-fourth of what TOEFL costs. All tests surveyed in Fig 1 except Pearson PTE are priced around the same price point.

Our solution (Fig 2) proposes to unify these varieties, allocating the available human budget intelligently to balance the reliability of the test with the cost to the test-taker. To the best of our knowledge, no existing systems target this continuum of utilizing both humans and AS raters. Providing this option would allow AS models to be deployed in more versatile scenarios, working in tandem with expert human raters to provide both *reliability* and *lower-cost* solutions. Increasing reliability helps to build trust in automatically scored exams, thus leading to broader adoption. Cost is a critical consideration to lower-income test-takers and those who need to take the test multiple times.

We define the problem and solution more formally as follows: given a set of responses to be scored, a target AS model, and an **expert human budget** (that is, the number of responses we can have scored by expert human raters), our goal is to *efficiently* sample responses to be scored by the expert. These expert-scored samples are then combined with automatically scored samples to *maximize* the overall system performance metric. We propose a novel Monte-Carlo sampling based reward sampling algorithm to efficiently sample responses to maximize the system performance.

Usually one or multiple amongst accuracy, Quadratic Weighted Kappa (QWK), or Cohen's kappa (Taghipour and Ng 2016; Zhao et al. 2017; Kumar et al. 2019; Grover et al. 2020; Singla et al. 2021a) are used in automatic scoring literature as they are robust measures of inter-rater reliability, a primary goal in Automated Scoring. A key point to be noted is that the reliability of the test (*i.e.* how consistently a test measures a characteristic) is measured on the **global** score (the aggregate of the responses) and lesser on the score on the individual responses. The global score determines admissions, interviews, and career growth, while per-item scores are used as indicators of particular skills.

While intuitively, we can say that there exists a monotonically increasing relationship between the reliability of the test on individual questions and the overall score, we show that it is more efficient to consider the global context instead of item-level context, while sampling responses for getting them double-scored by humans.

We establish strong baselines using **Uncertainty Sampling** (§3.2), an importance sampling formulation that samples using probability of being wrong output by the AS model. We propose **Reward Sampling** (§3.3), that samples based on the estimated *reward* of correcting a mistake.
We summarize our main contributions as follows:

- We propose to combine existing paradigms to integrate humans with Automated Scoring systems. Provided a budget indicating the number of responses that can be scored by human raters, we observe significant gains in accuracy and QWK using our proposed sampling model, **Reward Sampling** (§3.3). For instance, by using 40% human budget with an AS model with 64% accuracy, our sampling methodology can achieve an accuracy gain of 23% while random sampling leads to 14% and uncertainty sampling leads to 15%. To the best of our knowledge, this is the first time such a formulation has been considered in Automatic Scoring systems.

- We conduct experiments on various models differing in accuracy to show our algorithm's model agnostic nature (§3). We include results from models deployed in AS settings in the real world to crafted pseudo models. Averaging over these models, we observe 19.80% increase in accuracy and 25.60% increase in QWK when using reward sampling with 30% of the dataset as a human budget. The random sampling and uncertainty sampling baselines achieve 8.6% and 12.2% gains in accuracy, respectively.

- While augmenting the system's performance is an important goal, it is equally important to quantify this improvement, especially when deployed in the real world, where there are no labeled datasets to compare against and the consequences of misgrading, for both business and test takers, could be catastrophic. Thus, we also propose an algorithm to *estimate* the accuracy and QWK achieved, with statistical guarantees. (§3.4).

2 Related Work

Broadly, our paper covers two areas of research: Automatic Scoring and Sampling methods. Here we cover them briefly.

Automatic Scoring: The goal of an automatic scorer is to assess language competence of a candidate with an accuracy matching that of a human grader, but faster, with greater consistency and at a fraction of the cost (Malinin 2019; Yan, Rupp, and Foltz 2020). Almost all work in the automatic scoring domain has been to better model the scoring of essays and speech traits as a natural language processing task. The techniques have ranged from manually-engineered natural language features (Kumar et al. 2019; Dong and Zhang 2016) to LSTMs, memory networks (Zhao et al. 2017) and transformers (Singla et al. 2021a; Shah et al. 2021). There has also been some recent work in other facets of AS including adversarial testing (Ding et al. 2020; Kumar et al. 2020; Parekh et al. 2020), explainability (Kumar and Boulanger 2020), uncertainty estimation (Malinin 2019), off-topic detection (Malinin et al. 2016), evaluation metrics (Loukina et al. 2020), *etc.* The Linguaskill test escalates responses to be graded by humans when scores/confidence are outside manually set thresholds (Xu et al. 2020). To the best of our knowledge, there is no work on increasing the reliability of automatic scoring systems using sampling to bring humans into the loop. Most white papers from second language testing firms mention results on historical data as a measure of their reliability (Brenzel and Settles 2017; Pearson 2019). Due to continuous domain shift, historical results cannot be trusted for a model's future performance gains. Therefore, performance guarantees of AS models are essential to establish institutional trust in them. To fill this research gap, we propose reward sampling based on Monte Carlo sampling methods for measuring and increasing AS systems' reliability.

Monte-Carlo Sampling For Evaluation: There has been much work in improving automatic metrics using Monte-Carlo sampling methods in machine translation and natural language (NL) evaluation (Chaganty, Mussman, and Liang 2018; Hashimoto, Zhang, and Liang 2019; Wei and Jia 2021). They use statistical sampling methods like importance sampling and control variates to combine automatic NL evaluation with expensive human queries. To the best of our knowledge, we are the first to extend sampling techniques in the context of automatic scoring. We use them to combine relatively cheaper automatic scoring model results with expensive human expert scorers. Kang et al. (2020) use sampling for approximate selection queries. They combine cheap classifiers with expensive estimators to meet minimum precision or recall targets with guarantees. We extend their work to take the global context into account while estimating accuracy (§3.4).

3 System Overview

This section describes the components of the proposed solution, the intuition and reasoning behind the sampling mechanisms, and the algorithm for estimating the metrics with statistical guarantees. Given an Automated Scoring model, a dataset to be scored, and a human budget indicating the percentage of records we can provide to expert human raters for scoring, records are sampled making use of a pre-computed **human-machine agreement matrix** (to be described below). For the samples selected, we replace the predictions made by the AS model with the scores provided by the human raters and compute an estimate of the increase in accuracy and QWK with guarantees (Fig 2).

When considering sampling, the baseline approach is random sampling *i.e.* sampling with uniform probability for each record in the dataset. This is not a good allocation of resources, as when considering models of high quality, most samples will not provide any value. For example, with a model of 75% accuracy, random sampling would only provide value for $\sim$25% of samples, as the rest would have been correctly scored anyway. This motivates our search for a more efficient sampling mechanism, one that takes into account the probability of the model being wrong with respect

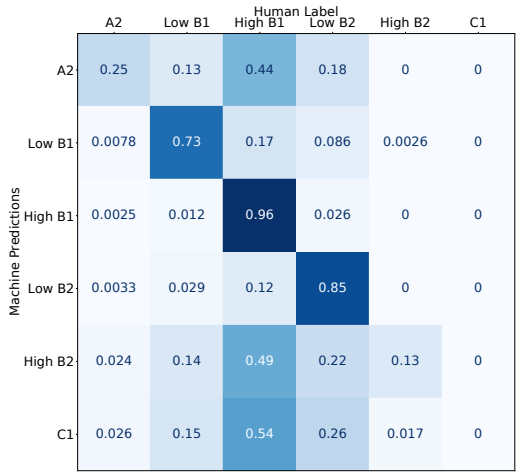


Figure 3: A sample human-machine agreement matrix on a CEFR aligned scoring scale. The rows indicate machine predictions, and each row is normalized to give the probability of the machine class matching the human labeled class.

to a human expert, and crucially, the reward that would be gained by correcting this mistake. We define the reward as the magnitude of the change in the *global* score that would occur when a *local* response is changed as a result of human correction of machine score (§3.3).

3.1 Human-Machine Agreement Matrix

The human-machine agreement matrix is a normalized confusion matrix of the model's predictions and the ground truth, precomputed on validation data or historical test data. Each entry indicates the probability of the class predicted by the machine aligning with the class labeled by the human. Fig 3 shows a sample human-machine agreement matrix where `m[Low B1][High B1] = 0.17` indicates the probability of the ground truth being `High B1` when the machine has predicted `Low B1`.

3.2 Uncertainty Sampling

The key idea behind uncertainty sampling is that the machine is not equally likely to be wrong across all prediction classes. Some scores may be assigned with much better accuracy than others. This idea is borne out by the human-machine agreement matrix as well, where the probabilities of a correct prediction are along the principal diagonal. We can see in Fig 3, `High B1, Low B2` are accurately predicted whereas `A2, High B2` predictions are likely to be wrong. Since the machine is likely making a wrong judgement when predicting these classes, it would be more efficient to sample more from the records where these predictions have been made and corrected using human labelers.

To quantify this, we formulate Uncertainty Sampling as vanilla importance sampling, where the *uncertainty* of each class is calculated using the cross-entropy function. Each row in the human-machine agreement matrix represents the probability distribution of the ground truth when that particular class has been predicted. The cross entropy of this distribution with the *ideal* distribution (one-hot encoding for that class) is calculated.

For *e.g.*, the distribution associated with `Low B1` in the matrix is $[0.0078, 0.73, 0.17, 0.086, 0.0026, 0]$. The cross entropy of this distribution with the ideal distribution ($[0, 1, 0, 0, 0, 0]$) for `Low B1` is calculated. In this way, we can quantify the "loss" associated with `Low B1`. Subsequently, every record is assigned a loss associated with the prediction made for that record, and this is normalized over the entire dataset to create a probability distribution. We draw a sample $s \sim U(D)$ without replacement from the uncertainty distribution over the dataset $U(D)$. The provided human budget indicates the number of samples to be drawn and the likelihood of a record being drawn corresponds to the uncertainty associated with the prediction class.

3.3 Reward Sampling

For single skill testing exams (for *e.g.*, one out of speaking, writing, reading) like the one by SLTI (2021) and LTI (2021), the test reliability and validity are measured over the complete test as opposed to individual prompts. While increasing accuracy on individual prompts (through sampling and subsequent human intervention) is a sure way of increasing the accuracy on the overall exam, it is more efficient to directly sample records which are *more likely to affect the overall result*, rather than simply sampling those which the machine is uncertain about. In uncertainty sampling, we sample records based on the likelihood of the prediction being wrong, but we do not consider whether *being right* would actually change the global score. This is the motivation behind reward sampling. Here we sample records which are more likely to generate a larger reward, *i.e.*, a change in the score at the global level. To this end, the expected reward E_R is calculated for each record in the dataset as:

$$E_R(d) = \sum_{c \in C} p(c \mid m) * reward(d, c) \quad \forall d \in D \qquad (1)$$

where d represents one record in the dataset D, c and m represent classes in the set of all classes C, $p(c \mid m)$ indicates the probability of the ground truth being c when machine has predicted m, and the reward function is denoted as $reward$. The expected reward encodes the reward gained by the ground truth being c when the machine has predicted m weighted by the *probability* of the same, summed over every class c. $p(c \mid m)$ is looked up from the human-machine agreement matrix and the output of the reward function is weighted by this probability.

The reward function calculates the reward gained by swapping the predicted class with a different class. The aggregate label for the candidate associated with d is calculated before and after the swap with a new class, and the reward is defined as the **absolute difference** between the two scores, which encodes the magnitude of the score change that would happen if the prediction class was changed from m to c. The absolute difference is considered because it is equally important if the new score is greater or lesser than the predicted score, thus incurring the same reward. If the prediction is an outlier compared to predictions on other responses of the same candidate, a large reward could be generated when

Model metrics (**global**)	Sampling Method	Accuracy Improvement Human Budget →					Kappa Improvement Human Budget →				
		10%	20%	40%	60%	80%	10%	20%	40%	60%	80%
BERT-Baseline acc - 0.66; kappa - 0.56	Random	0.69	0.72	0.78	0.84	0.93	0.59	0.63	0.7	0.78	0.9
	Uncertainty	0.7	0.73	0.8	0.86	0.93	0.59	0.62	0.71	0.8	0.9
	Reward	**0.74**	**0.82**	**0.88**	**0.91**	**0.95**	**0.64**	**0.76**	**0.84**	**0.88**	**0.94**
BERT-Two Stage acc - 0.69; kappa - 0.60	Random	0.73	0.75	0.81	0.87	0.93	0.65	0.68	0.75	0.83	0.91
	Uncertainty	0.72	0.75	0.82	0.87	0.94	0.63	0.66	0.74	0.82	0.92
	Reward	**0.79**	**0.86**	**0.91**	**0.92**	**0.96**	**0.72**	**0.81**	**0.87**	**0.9**	**0.94**
LSTM-Attn-Baseline acc - 0.64; kappa - 0.54	Random	0.67	0.71	0.78	0.85	0.93	0.58	0.63	0.71	0.79	0.9
	Uncertainty	0.68	0.72	0.79	0.88	0.93	0.57	0.6	0.69	0.82	0.9
	Reward	**0.73**	**0.78**	**0.87**	**0.92**	**0.96**	**0.62**	**0.71**	**0.83**	**0.89**	**0.95**
LSTM-Attn-Two Stage acc - 0.65; kappa - 0.57	Random	0.67	0.71	0.76	0.85	0.92	0.59	0.64	0.7	0.8	0.89
	Uncertainty	0.68	0.73	0.8	0.87	0.93	0.58	0.62	0.71	0.81	0.91
	Reward	**0.74**	**0.82**	**0.87**	**0.9**	**0.95**	**0.66**	**0.75**	**0.83**	**0.86**	**0.93**
Pseudo Model-0.75 acc - 0.72; kappa - 0.57	Random	0.74	0.76	0.8	0.86	0.93	0.62	0.64	0.72	0.81	0.9
	Uncertainty	**0.82**	**0.9**	**0.93**	**0.97**	**0.98**	**0.73**	**0.85**	**0.9**	**0.95**	**0.98**
	Reward	**0.81**	0.86	**0.92**	**0.96**	**0.98**	**0.73**	0.78	**0.89**	**0.94**	**0.97**

Table 1: Accuracy (**acc**) and Quadratic Weighted Kappa (**kappa**) for various models across multiple sampling methods and increasing percentages of the dataset as sample size. **Bold** indicates the best performing variant for each configuration.

changing predictions, making it a prime target for sampling. On the other hand, if changing the class to c does not change the final score, then a reward of 0 would be generated. With a zero reward, these records would not be sampled. Thus, to ensure that every record has a nonzero reward *i.e* a nonzero probability to be sampled, the reward is additively smoothed $E_R(d) = E_R(d) + \Delta$ where $\Delta = 0.001$. In this manner, an expected reward is calculated for each record in the dataset.

The sampling procedure proceeds similarly: the rewards are normalized to create a probability distribution over which a sample $s \sim E_R(D)$ is drawn. In using this sampling mechanism, we directly sample records that are most likely to provide us an improvement at the aggregate level, compared to indirectly improving the aggregate metrics when using uncertainty sampling.

3.4 Estimation with Guarantees

In high stakes testing scenarios, it is critical to ensure that the system does not fail catastrophically. For this reason, it is important to provide estimations of system metrics with guarantees. Kang et al. (2020) describes an algorithm that provides statistical guarantees on precision/recall on a subset of results returned from a dataset. More specifically, given a dataset, a precision/recall target value, sample size and a failure probability, the algorithm returns a result (a subset of the dataset) which meet the required precision/recall target with a *probabilistic guarantee*.

Our task is similar, but instead looks at providing guarantees on the accuracy/QWK of overall score on the entire dataset rather than just a dataset subset and individual samples. To provide these guarantees, we form confidence intervals over accuracy/QWK and take the lower bound.

The samples selected by reward and uncertainty sampling procedures are not a good fit for estimation as they have been taken with the purpose of correcting mistakes and improving reliability. This means that highly underconfident samples would be selected, thus leading to inaccurate performance estimates. Kang et al. (2020) show that importance sampling based on a model's confidence of prediction improve over uniform random sampling by providing a lower variance estimate. More specifically, they show that the squared confidence of the model minimizes the variance of the estimate. As we have not considered model confidence in our work, we take the following formulation as a proxy for confidence, applied over every *candidate* who wrote the test (not responses to individual questions):

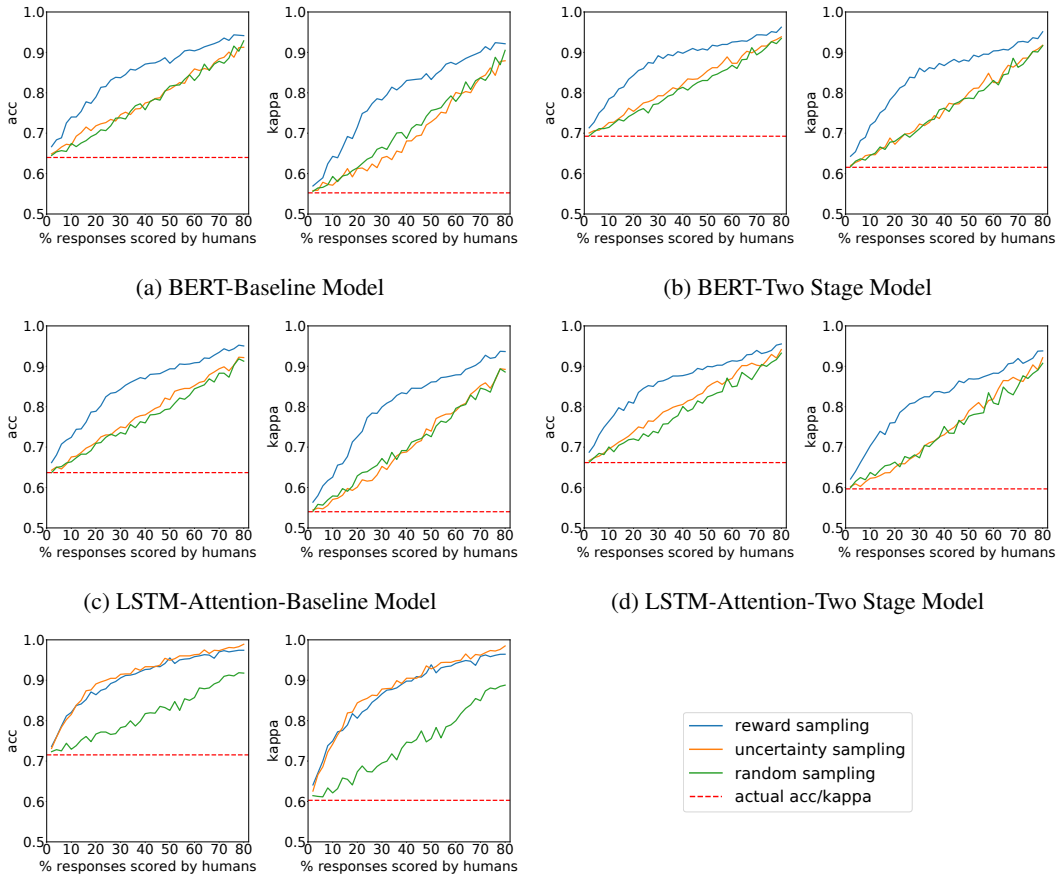
$$\zeta(t) = (1 - \sum_{i \in t} i[u]\,)^2 \quad \forall t \in T \tag{2}$$

where ζ represents the confidence associated with a test taker t in the set of all test-takers T, i represents individual responses of t and u represents the uncertainty. From uncertainty sampling, we have a normalized uncertainty associated with each response, this is aggregated over all responses of a candidate, subtracted from 1 and then squared to provide a confidence estimate. This confidence is normalized to create a probability distribution. A secondary smaller sample is taken over this distribution of *candidates*, effectively sampling all underlying responses of the candidate. Using the aggregated labels and predictions, the lower bound estimates of accuracy and kappa (McHugh 2012) are calculated.

4 Experiments

4.1 Dataset

To evaluate our method, we make use of data collected by Second Language Testing Inc. (SLTI) while conducting the Simulated Oral Proficiency (SOPI) Exam. The SOPI exam has been used since 1992 and studied extensively (Stansfield and Kenyon 1992, 1996). SOPI is used for interviews, university admissions, skill development and as a test in several online courses (SLTI 2021). SOPI offers psychometric advantages in terms of reliability and validity, particularly in standardized testing situations. The candidates in



(a) BERT-Baseline Model

(b) BERT-Two Stage Model

(c) LSTM-Attention-Baseline Model

(d) LSTM-Attention-Two Stage Model

(e) Pseudo Model with 0.75 accuracy (local)

(f) Legend

Figure 4: In each model, we show the change in accuracy (*left*) and quadratic weighted kappa (QWK) (*right*) after sampling with the sample size (human budget) shown on the x-axis. As can be seen, reward sampling outperforms both uncertainty sampling and random sampling baseline in each model.

the dataset are primarily Filipino high school graduates and above. A test-taker is presented with six prompts on their computer and their responses for each individual item are recorded. The prompts and the rubrics for evaluation follow the Central European Framework of Reference for Languages (CEFR) (Broeder and Martyniuk 2008) guidelines. The prompts difficulty varies from B1 to C1. A candidate receives both a prompt-level score and a global score calculated from the individual prompt-level scores. The SOPI dataset has eight question papers (forms) containing six prompts each, and each form was attempted by 7200 speakers on an average. Many other works have used the SLTI dataset for tasks including automated scoring and coherence modeling (Grover et al. 2020; Patil et al. 2020; Stansfield and Winke 2008; Singla et al. 2021a).

4.2 Experimental Setup

To demonstrate that the sampling methods described are model agnostic, we conduct experiments using multiple models of varying accuracy. We leverage speech scoring models from Singla et al. (2021a), making use of state-of-the-art models such as BERT and Bi-directional LSTMs, both baseline versions and conditioned on speaker information. In addition, we also run experiments on a pseudo model, described as follows. For a given accuracy, a pseudo model's predictions are generated by randomly changing $100 - acc\%$ of ground truth labels. For *e.g.*, the prediction of a pseudo model with 65% accuracy is the ground truth with 35% of labels randomly changed. Predictions, and hence accuracy, are generated at the *local* level, for each response whereas we are concerned about the metrics at the global level, which is typically lesser. The dataset is split into train and test sets, with the additional constraint that this split be done such that all responses of one candidate are contained in a set, and not split between the train and test sets. For our experiments, since we do not have a precomputed human-machine agreement matrix, we compute it using the training set and hold out the test set for verifying our proposed system. In addition, the aggregate dataset must also be calculated from each candidate's individual responses, calculating the candidate's global score from each of their responses.

The experiments were conducted with sample sizes upto

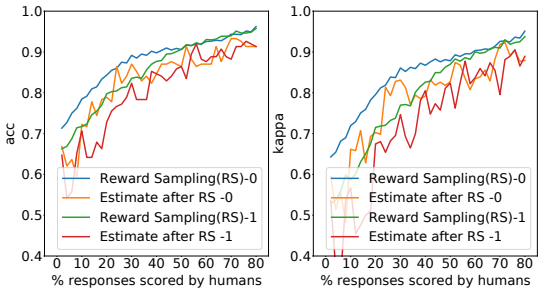


Figure 5: The metrics of two models [0 - BERT-TwoStage, 1 - LSTM-Baseline] have been presented. Results for other models are similar and are not shown for visual clarity.

80% of the dataset to observe the effect of sample size on the improvement in accuracy with respect to each sampling method. Records are sampled from the test set according to Reward Sampling, along with Random (uniform) Sampling and Uncertainty (importance) Sampling, our baselines. We replace the predictions of records in the sample with those of the ground truth, following which we recompute the aggregate dataset. This dataset is used to calculate the **system level metrics**, not just the model, but the combination of the model and the human in the loop. In estimating metrics, a secondary sample was taken. Empirically, we observed that a sample size of 200 was sufficient for stable estimations of a 95% confidence interval. We report our estimation on Reward Sampling when considering 80% of the dataset as human budget *i.e.* the most performant configuration.

5 Results

5.1 Improving Reliability

Table **??** presents the results of the experiments conducted across configurations of models, sampling methods, and human budget. Entries in **bold** indicate the best performing configuration, which is identical across nearly all models (Reward Sampling with the maximum sample size, 80% of the dataset). The models considered are BERT (baseline), BERT (two stage speaker conditioning), BD-LSTM with Attention (baseline), BD-LSTM with Attention (two stage speaker conditioning) and a pseudo model with $accuracy = 0.75$ at the local level. For all models, the dataset is aggregated, following which accuracy and QWK are calculated, giving us the values in the *Model metrics* column.

Fig 4 shows the change in both accuracy and QWK at the **global** level for various models. The changes are measured for increasing human budget *i.e.* percent of responses available to be scored by humans upto 80% of the dataset. For illustration, we consider the BERT-Baseline model. Firstly, we observe that random sampling shows minimal improvement (**3%**) over the actual accuracy when sampling 10% of the dataset to be scored by human raters. This is due to the model's accuracy (66% of all samples would have been predicted correctly anyway) and the remaining gains are further minimized by the aggregation process. Reward Sampling, on the other hand, shows an 8% gain in accuracy, more than twice the gains achieved by random sampling. Interestingly, Uncertainty Sampling shows similar gains to random sampling in all models except the pseudo model, where its performance is more in line with reward sampling. The predictions of the pseudo model are randomly generated, hence local gains translate well to global gains. This difference is likely the reason for the large gap in performance when considering uncertainty sampling on pseudo and real models.

Reward Sampling, where the reward that is gained by having a record rated by a human is also a sampling factor, shows significant gains across models as shown in Fig 4. We note that the gains provided by Reward Sampling decline compared to the baseline sampling methods with increasing sample sizes. Initially, Reward Sampling outperforms the other sampling methods with large gains, upto a sample size of $\sim$30%. Beyond this mark, the gains are no longer as significant and the other methods slowly catch up. This trend holds across all models, indicating that Reward Sampling shows maximal gains over baselines when sampling less than half of the dataset for human scoring.

5.2 Estimation with Guarantees

Fig 5 is a plot visualizing the metrics of the models when utilizing reward sampling and an estimate of the same. The sample size used for estimation remains constant and it is only the sample used for reward sampling that changes. After reward sampling, a sample based on the confidence distribution (§3.4) is drawn, and the 95% confidence interval for both accuracy and kappa is calculated. The lower bound is taken to provide a statistical guarantee that accuracy/QWK will only fall below the estimated values 5% of all runs.

6 Conclusion

Automatic Scoring (AS) helps assess the language competency of candidates with accuracy matching that of a human grader, but faster, with greater consistency and at a fraction of the cost. Existing systems either rely on double scoring, effectively scoring each sample by both human and AS system, or solely by an AS system. Although double scoring is more reliable, it is considerably more expensive. We develop novel, sample-efficient algorithms to target the spectrum of possible solutions in the middle of both extremes. We show that by using a relatively small human budget, we can improve and estimate performance with guarantees, thus increasing the reliability and trustworthiness of the system. We implement and evaluate our algorithms on real exam data, showing that they outperform naive baselines in all settings evaluated. These results indicate the promise of probabilistic algorithms to improve and estimate automatic scoring reliability with statistical guarantees.

As part of future research, we plan to work on even more sample efficient algorithms and incorporating trait scoring while sampling. Another possible research avenue where we can apply our algorithms is in test design. While right now test design involves linguistic validity assessment studies, it does not take into account the reliability of the final test built. Reliability of a test could be incorporated as another constraint easily through our modelling paradigm.

References

ACARA. 2018. ACARA News (January 2018). https://www.acara.edu.au/news-and-media/news-details?section=201801250300#201801250300. Accessed: 2021-12-14.

Brenzel, J.; and Settles, B. 2017. The Duolingo English Test—Design, Validity, and Value. *DET Whitepaper (Short)*.

Broeder, P.; and Martyniuk, W. 2008. Language education in Europe: The common European framework of reference. *Encyclopedia of language and education*, 209–226.

Chaganty, A. T.; Mussman, S.; and Liang, P. 2018. The price of debiasing automatic metrics in natural language evaluation. *arXiv preprint arXiv:1807.02202*.

Ding, Y.; Riordan, B.; Horbach, A.; Cahill, A.; and Zesch, T. 2020. Don't take "nswvtnvakgxpm" for an answer–The surprising vulnerability of automatic content scoring systems to adversarial input. In *Proceedings of the 28th International Conference on Computational Linguistics*, 882–892.

Dong, F.; and Zhang, Y. 2016. Automatic features for essay scoring–an empirical study. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, 1072–1077.

ETS. 2021. Find Test Centers and Dates for ETS TOEFL Exam. https://v2.ereg.ets.org/ereg/public/testcenter/availability/seats?_p=TEL. Accessed: 2021-12-14.

Grover, M. S.; Kumar, Y.; Sarin, S.; Vafaee, P.; Hama, M.; and Shah, R. R. 2020. Multi-modal automated speech scoring using attention fusion. *arXiv preprint arXiv:2005.08182*.

Hashimoto, T. B.; Zhang, H.; and Liang, P. 2019. Unifying Human and Statistical Evaluation for Natural Language Generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics.

Incorporated, M. 2017. PEG - The engine driving Automated Essay Scoring. https://utahcompose.com/sites/default/files/peg-Info-report.pdf. Accessed: 2021-12-14.

Kang, D.; Gan, E.; Bailis, P.; Hashimoto, T.; and Zaharia, M. 2020. Approximate selection with guarantees using proxies. *arXiv preprint arXiv:2004.00827*.

Kumar, V.; and Boulanger, D. 2020. Explainable Automated Essay Scoring: Deep Learning Really Has Pedagogical Value. In *Frontiers in Education*, volume 5, 186. Frontiers.

Kumar, Y.; Aggarwal, S.; Mahata, D.; Shah, R. R.; Kumaraguru, P.; and Zimmermann, R. 2019. Get IT Scored Using AutoSAS—An Automated System for Scoring Short Answers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9662–9669.

Kumar, Y.; Bhatia, M.; Kabra, A.; Li, J. J.; Jin, D.; and Shah, R. R. 2020. Calling out bluff: Attacking the robustness of automatic scoring systems with simple adversarial testing. *arXiv preprint arXiv:2007.06796*.

LaFlair, G. T.; and Settles, B. 2019. Duolingo English test: Technical manual. *Retrieved April*, 28: 2020.

Le, T. 2020. Testing & Educational Support in the US. https://my.ibisworld.com/us/en/industry/61171/key-statistics. Accessed: 2021-12-14.

Liu, J.; Xu, Y.; and Zhu, Y. 2019. Automated essay scoring based on two-stage learning. *arXiv preprint arXiv:1901.07744*.

Loukina, A.; Madnani, N.; Cahill, A.; Yao, L.; Johnson, M. S.; Riordan, B.; and McCaffrey, D. F. 2020. Using PRMSE to evaluate automated scoring systems in the presence of label noise. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 18–29.

LTI. 2021. Language Testing Inc. https://www.languagetesting.com/lti-information/general-test-descriptions. Accessed: 2021-12-14.

Malinin, A. 2019. *Uncertainty estimation in deep learning with application to spoken language assessment*. Ph.D. thesis, University of Cambridge.

Malinin, A.; Van Dalen, R.; Knill, K.; Wang, Y.; and Gales, M. 2016. Off-topic response detection for spontaneous spoken English assessment. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1075–1084.

McHugh, M. L. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3): 276–282.

O'Donnell, P. 2018. Computers are now grading essays on Ohio's state tests. https://www.cleveland.com/metro/2018/03/computers_are_now_grading_essays_on_ohios_state_tests_your_ch.html. Accessed: 2021-12-14.

Page, E. B. 1967. Statistical and linguistic strategies in the computer grading of essays. In *COLING 1967 Volume 1: Conference Internationale Sur Le Traitement Automatique Des Langues*.

Parekh, S.; Singla, Y. K.; Chen, C.; Li, J. J.; and Shah, R. R. 2020. My Teacher Thinks The World Is Flat! Interpreting Automatic Essay Scoring Mechanism. *arXiv preprint arXiv:2012.13872*.

Patil, R.; Singla, Y. K.; Shah, R. R.; Hama, M.; and Zimmermann, R. 2020. Towards Modelling Coherence in Spoken Discourse. *arXiv preprint arXiv:2101.00056*.

Pearson. 2019. Pearson Test of English Academic: Automated Scoring. https://assets.ctfassets.net/yqwtwibiobs4/26s58z1YI9J4oRtv0qo3mo/88121f3d60b5f4bc2e5d175974d52951/Pearson-Test-of-English-Academic-Automated-Scoring-White-Paper-May-2018.pdf.

Perelman, L.; Sobel, L.; Beckman, M.; and Jiang, D. 2014. The Basic Automatic B.S. Essay Language Generator (BABEL Generator). https://lesperelman.com/writing-assessment-robo-grading/babel-generator/. Accessed: 2021-12-14.

Service, E. T. 2020. Education Testing Service EIN 21-0634479. https://www.causeiq.com/organizations/educational-testing-service,210634479/. Accessed: 2021-12-14.

Shah, J.; Singla, Y. K.; Chen, C.; and Shah, R. R. 2021. What all do audio transformer models hear? Probing Acoustic Representations for Language Delivery and its Structure. *arXiv preprint arXiv:2101.00387*.

Singla, Y. K.; Gupta, A.; Bagga, S.; Chen, C.; Krishnamurthy, B.; and Shah, R. R. 2021a. Speaker-Conditioned Hierarchical Modeling for Automated Speech Scoring. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 1681–1691.

Singla, Y. K.; Parekh, S.; Singh, S.; Li, J. J.; Shah, R. R.; and Chen, C. 2021b. AES Systems Are Both Overstable And Oversensitive: Explaining Why And Proposing Defenses. *arXiv preprint arXiv:2109.11728*.

SLTI. 2021. Second Language Testing Inc. https://secondlanguagetesting.com/. Accessed: 2021-12-14.

Stansfield, C.; and Winke, P. 2008. Testing aptitude for second language learning. *Encyclopaedia of language and education, 2nd Edition: Language Testing and assessment*, 7: 81–94.

Stansfield, C. W.; and Kenyon, D. M. 1992. The development and validation of a simulated oral proficiency interview. *The Modern Language Journal*, 76(2): 129–141.

Stansfield, C. W.; and Kenyon, D. M. 1996. *Test Development Handbook: Simulated Oral Proficiency Interview,(SOPI)*. Center for Applied Linguistics.

Strauss, V. 2020. How much do big education nonprofits pay their bosses? Quite a bit, it turns out. https://www.washingtonpost.com/news/answer-sheet/wp/2015/09/30/how-much-do-big-education-nonprofits-pay-their-bosses-quite-a-bit-it-turns-out/. Accessed: 2021-12-14.

Taghipour, K.; and Ng, H. T. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, 1882–1891.

TechNavio. 2020. Global Higher Education Testing and Assessment Market 2020-2024. https://www.researchandmarkets.com/reports/5136950/global-higher-education-testing-and-assessment. Accessed: 2021-12-14.

Wei, J.; and Jia, R. 2021. The statistical advantage of automatic NLG metrics at the system level. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics.

Xu, J.; Brenchley, M.; Jones, E.; Pinnington, A.; Benjamin, T.; Knill, K.; Seal-Coon, G.; Robinson, M.; and Geranpayeh, A. 2020. Linguaskill: Building a validity argument for the Speaking test.

Yan, D.; Rupp, A. A.; and Foltz, P. W. 2020. *Handbook of automated scoring: Theory into practice*. CRC Press.

Yang, Y.; Buckendahl, C. W.; Juszkiewicz, P. J.; and Bhola, D. S. 2002. A review of strategies for validating computer-automated scoring. *Applied Measurement in Education*, 15(4): 391–412.

Zechner, K.; Higgins, D.; Xi, X.; and Williamson, D. M. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, 51(10): 883–895.

Zhao, S.; Zhang, Y.; Xiong, X.; Botelho, A.; and Heffernan, N. 2017. A memory-augmented neural model for automated grading. In *Proceedings of the fourth (2017) ACM conference on learning@Scale*, 189–192.