# An apprenticeship model for human and AI collaborative essay grading

**Alok Baikadi**
alok.baikadi@pearson.com
Pearson

**Lee Becker**
leek.becker@pearson.com
Pearson

**Jill Budden**
jill.budden@pearson.com
Pearson

**Peter Foltz**
peter.foltz@pearson.com
Pearson

**Andrew Gorman**
andrew.gorman@pearson.com
Pearson

**Scott Hellman**
scott.hellman@pearson.com
Pearson

**William Murray**
william.murray@pearson.com
Pearson

**Mark Rosenstein**
mark.rosenstein@pearson.com
Pearson

## CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**.

## KEYWORDS

Human-ai collaboration; Intelligent user interfaces; Machine learning; Active learning; Automated essay grading

**ACM Reference Format:**
Alok Baikadi, Lee Becker, Jill Budden, Peter Foltz, Andrew Gorman, Scott Hellman, William Murray, and Mark Rosenstein. 2019. An apprenticeship model for human and AI collaborative essay grading. In *Joint Proceedings of the ACM IUI 2019 Workshops, Los Angeles, USA, March 20, 2019* , 6 pages.

## 1 INTRODUCTION

Across a range of domains, humans and complex algorithms embedded in systems are collaborating and coordinating action. Tasks are shared because the combined system can be more capable than either agent acting alone. Such systems with shared autonomy raise important research questions in how to design these joint systems to best utilize each of the actors capabilities for optimal performance while addressing important safety, legal and ethical issues.

Our work investigates students developing their writing ability throughout their educational career, since writing proficiency is a critical life and career competency. Writing is a skill that develops through practice and feedback. However, the massive effort required of instructors in providing feedback on drafts and scoring final versions is a limiting factor in assigning essays and short answer problems in their classes.

Over the last 20 years, automated scoring of student writing through the use of natural language processing, machine learning and artificial intelligence techniques coupled with human-scored training sets has in many applications achieved performance comparable to that of humans [19]. In high stakes testing, millions of students have had their writing automatically scored, with the prompts, rubrics and human scoring for training and validating model performance implemented in tightly controlled frameworks. Our work addresses the question of how to move this technology to a wider audience and adding formative assessment to have a broader impact in helping students learn to write.

Our goal is to lower the barriers that limit an instructor's ability to assign essays in their courses. Our approach is to develop a system in which instructors develop prompts appropriate for their course and, by scoring a subset of their student responses, are able to turn on an automated system to score the rest of the responses. These prompts and the ability to automatically score them become a resource that instructors can reuse and even share. The critical issue is that while the instructor is an expert in their domain, they are likely not an expert in either assessment or the machine learning techniques that make automated scoring possible.

We have piloted a prototype system in a large introductory psychology course at a large university. This pilot explored the issues of 1) transferring scoring expertise from an instructor to an automated system and 2) using an automated system to provide feedback to the instructor about the quality of the current state of its scoring. In most end user applications of AI, the user is only exposed to decisions or behavior from some unseen, unknown model, and training of the machine learning mechanism is either hidden or taken for granted by the user. Exposing machine learning flows to users unversed in the notions of model performance and evaluation raises interesting design questions around user trust, system transparency, and managing expectations.

*IUI Workshops'19, March 20, 2019, Los Angeles, USA*

We discuss the approach used for the pilot and some of the issues that emerged. We then show how adopting a metaphor of apprenticeship clarifies the communication between the instructor and the AI assistant. In this paper we discuss the issues of shared autonomy that arise in such a system and the issues we have seen, in defining the task, in making results mutually interpretable to the instructor and the machine, and in designing user interfaces that makes more transparent all the various factors that are required for making correct decisions on how the task should proceed.

## 2 RELIABLE SCORING

Current high stakes scoring implementations attempt to achieve reliability through various means: the use of rubrics to specify the results that must be evidenced at each score point [9, 20]; anchor papers, which are example, actual essays selected to indicate typical and boundary score point answers [15]; and supervised scorer training, which often includes practice scoring exercises and required levels of performance on example essays. Yet despite these rigorous preparations, raters still disagree. Psychometricians have developed methods to detect subtle biases in raters referred to as rater effects [23]. These techniques can allow performance monitoring during scoring and detection after scoring.

In complex tasks, such as writing an essay, even with well-trained scorers without rater effects there will be an expected level of disagreement over the score of individual essays. A number of studies have found that in well-designed prompts and rubrics with well-trained scorers, the expected range of adjacent agreement, i.e., scores within 1 point, is 80-99% and correlations in the 0.7 to 0.8 range (Brown et al. [2] provide a summary of research and standards in this area).

Human scoring is time consuming, expensive and limits the immediacy of feedback that can be provided to the student. Page described the first system to automatically score essays based on analysis of a fairly limited set of features from the essay [16]. Present day automated systems score millions of student essays in both high stakes and formative engagements with performance levels at or above human scorers (Shermis and Burstein have co-edited comprehensive summaries on the subject [18, 19]). These automated scoring systems are typically based on supervised machine learning, where the system is trained on a set of student essays and human scores. The system derives a set of features for each essay and learns to infer the human score from the features. A sample of essays, typically on the order of 300 to 500, are scored by human scorers, and then used to train the automated system. Performance of the automated system is compared to the performance of two human raters and, if found acceptable, the automated system then scores the remaining essays.

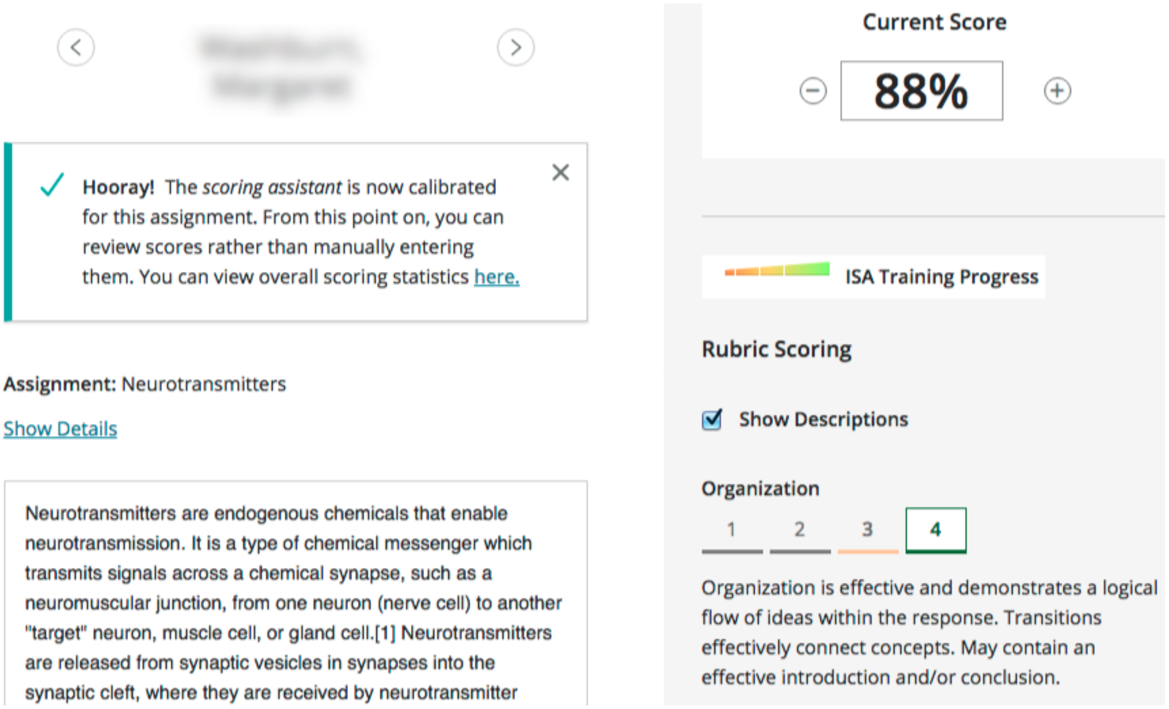
In the six years since the most current survey of the automated scoring field [19] developments in machine learning have impacted the modeling choices in both research and in commercial systems. These techniques include hierarchical classification [14], correlated linear regression [17] and various neural net deeplearning approaches [7, 21] and many others. In addition, some commercial systems have described their modeling subsystems, e.g. [4].

As we move away from high stakes scoring with precisely trained models to formative scoring with instructor-trained models (and the use of automated scoring in the classroom), the burden of generating reliable scores to train the automated system now falls on the instructor. For the automated system to reliably score, the instructor must score a sufficiently large number of essays to capture the variability of student responses and do so in a sufficiently reliable fashion to allow the regularities of scoring to be learnable by machine learning techniques. In the system we have developed, where the system learns from an instructor's scoring behavior, the instructor only need score enough essays to build a performant model. The hurdle is that, as the instructor scores, the system needs to provide feedback to the instructor as the instructor scores on how well the current model is performing. In an intelligible manner, the system must update the instructor on its progress. The AI system must continually provide information to the instructor to allow the instructor to make an informed decision about the quality of the automated scoring and determine when it is justifiable to turn scoring over to the automated system.

## 3 SYSTEM DESCRIPTION

We have developed a prototype system which allows instructors to assign writing to their students and participate in AI-assisted grading workflows. The system allows an instructor to create an account, invite students to join a course, and assign writing within a course. The prototype reported on here is an intermediate step toward enabling instructors to write and have their own prompts automatically scored. This step allows us to test the user interface and methods for sharing the task between the instructor and the system. This system learns to modify the scoring of existing, already-modeled prompts to more closely represent an instructor's scoring. In this current iteration, instructors select from a list of available writing prompts, each of which contains a short description, a rubric against which to grade student submissions, and a currently existing automated grading model. Once the prompt has been assigned, students are able to draft and submit their responses.

The collection of student responses goes through an active learning preprocessing step to calculate a recommended ordering for the instructor to grade essays. Active learning is typically employed to reduce human annotation effort, and in our system we use it to minimize the number of human-graded submissions needed for reliable modeling. Within



✓ **Hooray!** The *scoring assistant* is now calibrated for this assignment. From this point on, you can review scores rather than manually entering them. You can view overall scoring statistics here.

**Assignment:** Neurotransmitters

Show Details

Neurotransmitters are endogenous chemicals that enable neurotransmission. It is a type of chemical messenger which transmits signals across a chemical synapse, such as a neuromuscular junction, from one neuron (nerve cell) to another "target" neuron, muscle cell, or gland cell.[1] Neurotransmitters are released from synaptic vesicles in synapses into the synaptic cleft, where they are received by neurotransmitter

**Current Score**

88%

ISA Training Progress

**Rubric Scoring**

☑ Show Descriptions

**Organization**

Organization is effective and demonstrates a logical flow of ideas within the response. Transitions effectively connect concepts. May contain an effective introduction and/or conclusion.

**Figure 1: The instructor's grading interface used in our pilot study. Upper Left: AI readiness notification message. Lower Left: Essay prompt details link and student's essay response. Right: Total score, AI training progress bar, and rubric scoring controls with score of 4 for trait Organization.**

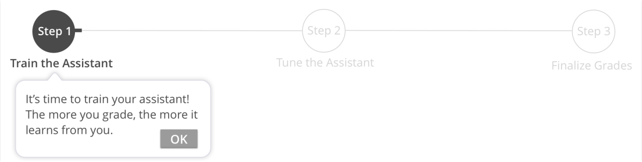
the instructor interface this is simply the list of submissions to grade sorted by the active learning order. In our current implementation, we use the Kennard-Stone algorithm [10]. Kennard-Stone attempts to select submissions in a manner that uniformly covers the feature space by iteratively selecting the submission with the maximum minimum distance to all previously selected submissions. We use baseline automated scores as our feature space so that the human grader will see approximately the same number of submissions at each score point, despite very low and very high scoring submissions being rare. In other natural language active learning tasks, biasing the active learner in favor of low-frequency classes has been found to work well [8, 22], and Kennard-Stone has been found to perform well for automated grading in particular [6].

As the instructor scores submissions, the system begins the modeling phase. In the modeling phase, the machine learning system is trained to mimic the instructors grades, and its performance is evaluated. Once the system determines the evaluation is acceptable, the instructors are signaled that the training is complete, and they are able to view the automated grades and make adjustments as needed. If the instructor corrects the grade, the system refines the model using the newly graded submissions.

As a first step to understand how instructors interact with AI systems, we decided to not allow instructors direct access to a highly complex large parameter space machine learning model. Instead, instructors assigned prompts for which there already existed a fully trained machine learning model. Our implementation uses logistic regression to learn a model that modifies the pre-trained automated scores to better match the instructor's scoring. The system learned the two parameters of a logistic regression model to estimate the instructor's scores based on the instructor's scores on responses and the scores from the existing model. The logistic regression functions as a transformation over the pre-trained model scores by adjusting the distance between score points to more closely match the instructor's scoring behavior. By learning a transformation over the pre-trained model, we are able to leverage the accuracy of the existing model, while allowing instructors to adjust the scoring needs to suit their classroom needs.

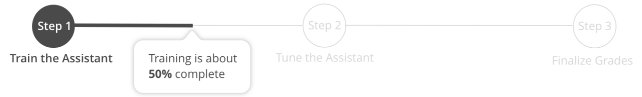
## 4 PILOT STUDY

We conducted a pilot with nine instructors and teaching assistants for an Introductory Psychology course at a large university. The participants completed an initial training session where they were presented with an overview of the interface,



It's time to train your assistant! The more you grade, the more it learns from you.

**Figure 2: Step 1a - Model training (apprenticeship *modeling*): UI indicates instructor must explicitly train the AI grading assistant. This is the initial messaging.**



Training is about **50%** complete

**Figure 3: Step 1b - Model training (apprenticeship *modeling*): UI indicates current progress in the training phase.**

and were encouraged to practice on a small set of student writing before the start of the pilot. Over the course of the semester, the participants were asked to use our system to grade 100 student submissions for each of nine writing prompts. Student submissions were sampled from the participating instructors' courses and were anonymized. Prompts were assigned in sets of three, and then made available to the participants to score.

Participants received prompts to grade in three phases, each consisting of three prompts. Upon logging in, the participants were able to begin grading by selecting any of the three available prompts to begin grading. The prompt was presented on a summary screen (Figure 1). The summary screen also presents the participants with the suggested order in which to grade the submissions (generated by active learning).

In addition to this ordering, the submissions were further divided into two sets: One that must be graded by the human rater, and one that could have AI feedback. As participants graded the first set, a progress bar indicated how close the model was to being trained. They could refer to the summary screen to evaluate their progress at any time. Once the predetermined threshold was reached, the participants received a "Hooray" message, as in Figure 1, and they were then able to review the output of the automated scoring model, adjusted by the logistic regression. They would then review several autoscored responses, and adjust the scores as needed. The regression model would be recalculated after each further adjustment. Once the participants were satisfied with the performance of the model, they were able to finalize the scoring and fix the grades for the rest of the submissions to that prompt.

## 5 APPRENTICESHIP MODEL OF TRAINING

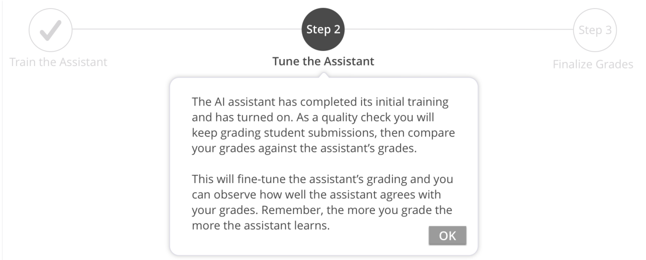
The system used in the pilot employed a progress bar with an alert message to communicate the current level of scoring performance to the participants. The interface and interaction implicitly encouraged instructors to regard the system as a tool and to think of system state as bimodal – untrained or trained. The disadvantage of this approach is that it encouraged instructors to infer that after the transition from untrained to trained the tool's performance matched their own, but technically it meant that a fuzzy threshold had been passed but further monitoring and feedback of automated scoring were still required.

This mismatch likely caused participants to be less vigilant, while survey results indicated participants felt disappointed when they had to correct the nascent automatic scoring. The message "Hooray! The scoring assistant is now calibrated . . . " and the green color of the progress bar implicitly set incorrect expectations and discouraged participants from carrying out further review and revision of scores, other than minimal tests to satisfy themselves that the model was performant. Additionally, during pre-pilot instruction, we suggested that the participants review approximately twenty submissions after the autoscoring model was enabled. Participants rarely strayed from these guidelines, reviewing approximately twenty submissions on average. In post-surveys, participant responses indicated that they did not have a strong sense of when to stop reviewing. Many would grade until the automated scores for a single essay matched their expectations. Analyses of behavioral data, survey results and users' feedback motivated us to reevaluate our user experience design to better scaffold the user through the process of training and to better communicate the expected quality of the automated scoring model.

While apprenticeship has been a model of human skill building for millennia, Lave was among the first to study and describe it as a formal mode of learning [12]. Collins et al. further generalized Lave's observation into what we refer to as an apprenticeship model of training [5]. This pedagogy-oriented paradigm consists of multiple phases, where the first three are relevant for our application: modeling, coaching, and fading. In modeling the apprentice (learner) "repeatedly observes the expert performing the target process". During coaching, the apprentice "attempts to execute the process while the expert provides guidance and scaffolds feedback and instruction". Lastly, in fading the expert provides less feedback and eventually ascertains the apprentice's mastery.

This paradigm has been employed for computer supported collaborative learning (CSCL) and intelligent tutoring systems (ITS) where the system regards the user as an apprentice to help them develop new skills [3, 11, 13].

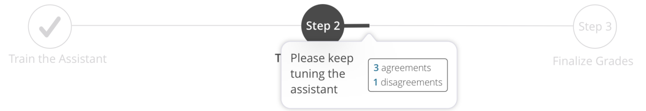
The apprenticeship model provides a useful metaphor for aligning our system's three stages of data gathering and application with an accessible, real-world process. Our system swaps the human-computer relationship typical of ITS and



The AI assistant has completed its initial training and has turned on. As a quality check you will keep grading student submissions, then compare your grades against the assistant's grades.

This will fine-tune the assistant's grading and you can observe how well the assistant agrees with your grades. Remember, the more you grade the more the assistant learns.

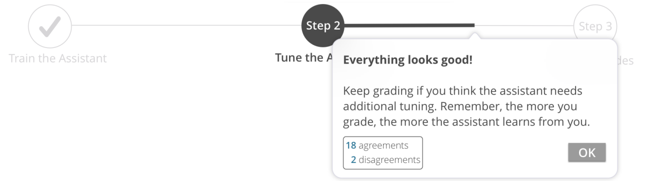
**Figure 4: Step 2a - Model tuning/validation (apprenticeship *coaching*): Messaging informs instructor AI grading assistant is ready to begin scoring essays.**



Please keep tuning the assistant

3 agreements
1 disagreements

**Figure 5: Step 2b - Model tuning/validation (apprenticeship *coaching*): Evaluation metrics inform instructor of AI grading assistant's performance and encourages the instructor to continue grading.**



Everything looks good!

Keep grading if you think the assistant needs additional tuning. Remember, the more you grade, the more the assistant learns from you.

18 agreements
2 disagreements

**Figure 6: Step 2c - Model tuning/validation (apprenticeship *coaching*): Messaging informs instructor AI grading assistant might be able to take over scoring essays.**

instead considers the user the expert and the AI-assistant persona the apprentice. By adopting this framework of model training as task *modeling*; model tuning and validation as *coaching* and model acceptance as *fading*, we help the user to better understand the expected interactions and responsibilities.

Our coupling of apprenticeship with machine learning is distinct from the use of apprenticeship in reinforcement learning [1], which does not have an interactive human element.

## 6 REDESIGNED USER INTERFACE AND FLOW

Based on the feedback from our initial pilot, in our new three-phase apprenticeship approach, we encourage instructors to view the process of training the automated scoring system as an apprenticeship. In this view, they can reasonably expect the AI-assistant to continue to learn even after it starts grading. The instructor now expects mistakes to continue, but in diminishing number and severity over time. With this approach minor mistakes are less likely to damage trust in the system.

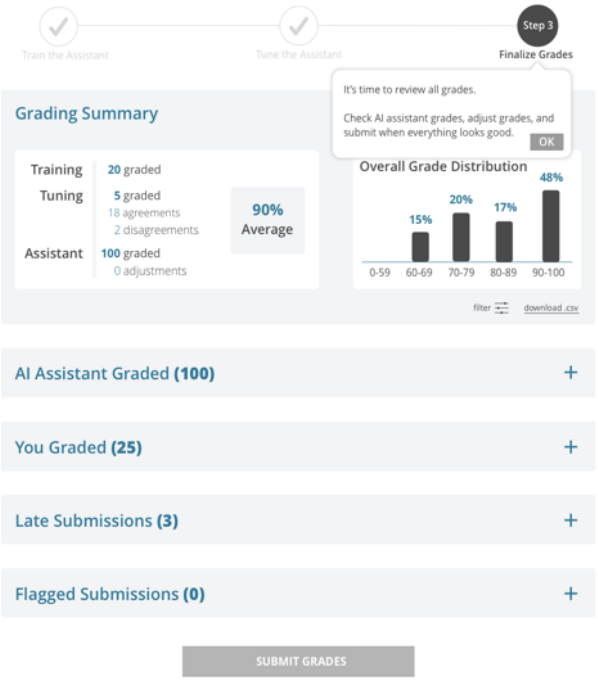
In the redesigned UI, at the top of each screen a large circle indicates the current location in the process, with messages updating to keep the instructor informed of progress within a given phase. At the beginning of the apprentice *modeling* phase (Figure 2), instructors are encouraged to score essays to help train the AI grading assistant. As they score more essays they see progress updates as shown in Figure 3. When the AI grading assistant achieves the ability to initially begin scoring, instructors move into the *coaching* phase (Figure 4).

In this phase the instructor scores and then compares their score to the AI-assistant (Figure 5). This phase ends when the system gains sufficient confidence in the model's performance (e.g., 0.7 to 0.8, values similar to human inter-rater reliability). This is reflected in the instructor's view by showing the instructor's corrections diminish (Figure 6). Once in the *fading* phase (Figure 7), the instructor passes scoring to the AI-assistant, but still retains the ability to review the AI-assistant's scores. Messages reinforce the relationship between additional scoring and performance, making the apprentice-relationship of the assistant (e.g., "...the more you grade, the more the assistant learns from you") more transparent. The level of the assistant's learning is indicated by the number and percentage of agreements compared to disagreements with the instructor's grades, and by the progress bar. The progress indicated by the bar follows the number of essays scored, but can accelerate as model performance improves.

## 7 CONCLUSIONS AND FUTURE WORK

As more people interact with systems based on sophisticated, often opaque algorithms, it becomes ever more critical to develop common languages and appropriate metaphors to allow communication and common understanding. Often, as these systems move to increasingly common use, a more refined understanding of how the machine learning component is trained and what its limitations are, becomes lost. In our first pilot we adopted a quite reasonable view of training an automated scoring system as a tool. Our first set of instructors internalized this model with unexpected consequences both for their performance on the task and their satisfaction with completing the task. In moving to the apprentice model, we believe we have found a metaphor that ameliorates some of these issues.

Our next steps include conducting pilots with this new metaphor and a UI/UX that supports it. We have begun to think more broadly about the complex relationships between clever systems and equally clever people, both of which have large blind spots. The instructors know the domain area but may have less experience in the type of reliable scoring required to train an automated scoring model. The AI system



It's time to review all grades.

Check AI assistant grades, adjust grades, and submit when everything looks good.

Grading Summary

Training 20 graded

Tuning 5 graded
18 agreements
2 disagreements

Assistant 100 graded
0 adjustments

90%
Average

AI Assistant Graded **(100)**

You Graded **(25)**

Late Submissions **(3)**

Flagged Submissions **(0)**

**Figure 7: Step 3 - Model acceptance (apprenticeship *fading*): Instructor has handed over control to AI grading assistant and is encouraged to review scores on remaining essays.**

embodies knowledge about scoring that can be used to scaffold the instructor's scoring, but at the same time is an apprentice to how the instructor wants the prompt evaluated. How to share the task and how the two agents should communicate are interesting, open questions. These questions will become even more relevant as we will begin testing the complete system which will now include instructors authoring prompts and replacing logistic regression with a full modeling pipeline.

## REFERENCES

[1] P. Abbeel and A. Ng. 2004. Apprenticeship Learning via Inverse Reinforcement Learning. In *Proceedings of the 21st International Conference on Machine Learning*.

[2] Gavin TL Brown, Kath Glasswell, and Don Harland. [n. d.]. Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing writing* 9, 2 ([n. d.]), 105–121.

[3] John Seely Brown, R. R Burton, and A. G. Bell. 1975. SOPHIE: A Step toward creating a reactive learning environment. *International Journal of Man-Machine Studies* 7, 5 (Sept. 1975), 675–696.

[4] Jing Chen, James Fife, H, Issac I. Bejar, and André A. Rupp. 2016. Building e-rater® Scoring Models Using Machine Learning Methods. *ETS Research Report Series 2016.1* (2016), 1–12.

[5] A. Collins, J. S. Brown, and S. E. Newman. 1987. *Cognitive apprenticeship: Teaching the craft of reading, writing and mathematics*. Technical 403. Centre for the Study of Reading, University of Illinois, BBN Laboratories, Cambridge, MA.

[6] Nicholas Dronen, Peter W. Foltz, and Kyle Habermehl. 2015. Effective Sampling for Large-scale Automated Writing Evaluation Systems. *Proceedings of the Second (2015) ACM Conference on Learning @ Scale* (2015), 3–10. https://doi.org/10.1145/2724660.2724661

[7] Youmna Farag, Helen Yannakoudakis, and Ted Briscoe. 2018. Neural automated essay scoring and coherence modeling for adversarially crafted input. *arXiv preprint arXiv:1804.06898* (2018).

[8] Andrea Horbach and Alexis Palmer. 2016. Investigating Active Learning for Short-Answer Scoring. In *BEA@ NAACL-HLT*. 301–311.

[9] Anders Jonsson and Gunilla Svingby. 2007. The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review* 2, 2 (2007), 130–144.

[10] Ronald W. Kennard and Larry A. Stone. 1969. Computer Aided Design of Experiments. *Technometrics: A Journal of Statistics for the Physical, Chemical, and Engineering Sciences* 11, 1 (1969), 137–48.

[11] S. P. Lajoie and A. M. Lesgold. 1992. Dynamic assessment of proficiency for solving procedural knowledge tasks. *Educational Psychologist* 27 (1992), 365–384.

[12] J. Lave. [n. d.]. *Tailored learning: Education and everyday practice among craftsmen in West Africa*. Technical Report.

[13] A. Lesgold, G. Eggan, and G. Rao. 1992. Possibilities for assessment using computer-based apprenticeship environments. *Cognitive approaches to automated instruction W. Regian & V. Shute (Eds.)* (1992), 49–80.

[14] Danielle S. McNamara, Scott A. Corssley, Rod D. Roscoe, Laura K. Allen, and Jianmin Dai. 2015. A hierarchical classification approach to automated essay scoring. *Assessing Writing* 23 (2015), 35–59.

[15] Miles Myers. 1980. *A procedure for writing assessment and holistic scoring*. National Council of Teachers of English, Urbana, IL.

[16] Ellis B. Page. 1967. Statistical and linguistic strategies in the computer grading of essays. *Coling 1967: Conférence Internationale sur le Traitement Automatique des Langues, Grenoble, France* (1967).

[17] Peter Phandi, Kian Ming A. Chai, and Hwee Tou Ng. 2015. Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

[18] Mark D. Shermis and Jill C. Burstein (Eds.). 2003. *Automated essay scoring: A cross-disciplinary perspective*. Lawrence Erlbaum Associates, Inc., Mahway, NJ.

[19] Mark D. Shermis and Jill C. Burstein (Eds.). 2013. *Handbook of automated essay evaluation: Current applications and new directions*. Routledge, New York.

[20] Dannelle D. Stevens and Antonia J. Levi. 2013. *Introduction to Rubrics: An Assessment Tool to Save Grading Time, Convey Effective Feedback, and Promote Student Learning*. Stylus Publishing, LLC.

[21] Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

[22] Katrin Tomanek and Udo Hahn. 2009. Reducing Class Imbalance During Active Learning for Named Entity Annotation. In *Proceedings of the Fifth International Conference on Knowledge Capture (K-CAP '09)*. ACM, New York, NY, USA, 105–112. https://doi.org/10.1145/1597735.1597754

[23] Edward W. Wolfe. 2004. Identifying rater effects using latent trait models. *Psychology Science* 46 (2004), 35–51.