Automated Scoring: Beyond Natural Language Processing

Nitin Madnani Aoife Cahill

Educational Testing Service
Princeton, NJ, 08541 USA
{nmadnani,acahill}@ets.org

Abstract

In this position paper, we argue that building operational automated scoring systems is a task that has disciplinary complexity above and beyond standard competitive shared tasks which usually involve applying the latest machine learning techniques to publicly available data in order to obtain the best accuracy. Automated scoring systems warrant significant cross-discipline collaboration of which natural language processing and machine learning are just two of *many* important components. Such systems have multiple stakeholders with different but valid perspectives that can often times be at odds with each other. Our position is that it is essential for us as NLP researchers to understand and incorporate these perspectives into our research and work towards a mutually satisfactory solution in order to build automated scoring systems that are accurate, fair, unbiased, and useful.

1 What is Automated Scoring?

Automated scoring is an NLP application usually deployed in the educational domain. It involves automatically analyzing a student's response to a question and generating either (a) a score in order to assess the student's knowledge and/or other skills and/or (b) some actionable feedback on how the student can improve the response (Page, 1966; Burstein et al., 1998; Burstein et al., 2004; Zechner et al., 2009; Bernstein et al., 2010). It is considered an NLP application since typically the core technology behind the automated analysis of the student response enlists NLP techniques. The student responses can include essays, short answers, or spoken responses and the two most common kinds of automated scoring are the automated evaluation of writing quality and content knowledge. Both the scores and feedback are usually based on linguistic characteristics of the responses including but not limited to:

(i) Lower-level errors in the response (e.g., pronunciation errors in spoken responses and grammatical/spelling errors in written responses),

(ii) The discourse structure and/or organization of the response,

(iii) Relevance of the response to the question that was asked.

2 Motivation

Over the last few years, there has been a significant increase in the number of NLP conference and workshop publications on the task of automated scoring of student responses (Burrows et al., 2015; Zesch et al., 2015; Sultan et al., 2016). Much of this increase in interest stems from the public availability of fairly large datasets containing scored human responses as part of shared tasks and public contests (the ASAP[1] and ASAP2[2] Kaggle shared tasks, the Powergrading dataset (Basu et al., 2013), and the SemEval

[1] https://www.kaggle.com/c/asap-aes

[2] https://www.kaggle.com/c/asap-sas

Proceedings of the 27th International Conference on Computational Linguistics, pages 1099–1109
Santa Fe, New Mexico, USA, August 20-26, 2018.

2013 Shared Task (Dzikovska et al., 2013)). Although this increase in interest is well-motivated and has led to useful technical advances, it has also propagated the impression that the task of automated scoring can be a self-contained, moderately simple machine learning task requiring "only" sophisticated feature engineering (Somasundaran et al., 2015; Ghosh et al., 2016) or, more recently, complex neural network architectures (Taghipour and Ng, 2016; Dong et al., 2017; Riordan et al., 2017; Tay et al., 2017). Our motivation for writing this paper is to highlight the differences between automated scoring as a shared task and automated scoring as an area of NLP research that serves the end goal of building and deploying accurate and unbiased scoring models for use in actual assessments or classrooms. For the latter, we argue that it is essential for NLP researchers to interact with and, indeed, collaborate with additional parties who are impacted by and contribute to automated scoring.[3]

To be clear, the issues we highlight are not simply a result of "operationalizing" NLP research. Instead, we claim that in order to make meaningful contributions in the area of automated scoring, these issues must be thought about and discussed as part of the NLP research and development process from the start. Our aim is not to strike a prescriptive tone. We acknowledge that despite the best of intentions, practical and business considerations can sometimes impact the extent to which such desiderata can be incorporated into the research process. However, our position is that these are extremely important problems to solve and even limited engagement with and progress towards solving these problems constitutes an important step. As more and more educational institutions and start-ups opt into automated scoring of educational assessments — which has the potential to impact the lives and livelihoods of members of the public — it is incumbent upon us as NLP researchers to strive towards fair, accountable, and transparent utilization of our research (Barocas et al., 2017).

3 Related Work

Many of the issues we refer to in the previous section are commonly discussed in the related fields of psychometric measurement and educational research. For example, Table 1 lists some of the standards in the 2014 Standards for Educational and Psychological Testing which are relevant for automated scoring. These standards help ensure fair and valid tests and provide guidelines for all aspects of assessment development. In addition, these communities have also been working towards best practices for the use of automated scoring (Clauser et al., 2002; Yang et al., 2002; Williamson et al., 2012; Bridgeman et al., 2012; Bejar et al., 2016). However, since there are limited opportunities for that community to interact with the NLP community, our goal is to bring them to the attention of the mainstream NLP audience as a first step towards the cross-discipline collaboration that we argue for in this paper.

In addition, there have also been efforts in the NLP community itself to highlight how NLP research on automated scoring can be situated in cross-disciplinary contexts (e.g. the workshops on Innovative Use of NLP for Building Educational Applications (BEA) and the workshops on NLP Techniques for Educational Applications (NLPTEA)) (Napoles and Callison-Burch, 2015; Wilson and Martin, 2015; Burstein et al., 2016; Beigman Klebanov et al., 2016; Zalmout et al., 2016; Lugini and Litman, 2017; Pado, 2017; Yaneva et al., 2017). However, the topics of these workshops tend not to always make it into the mainstream NLP conference discussions. Yet, at the same time there has been an increasing interest in automated scoring techniques at main NLP conferences. It is important not to lose the link between developing new and improved NLP techniques for automated scoring and the contexts in which they are generally applied.

4 Perspectives on Automated Scoring

Before we describe our position, we describe the various entities who are likely to be affected by automated scoring systems, and their corresponding perspectives on the use of automated scoring. All of these entities have a stake in making sure that the automated scoring system performs in line with their own expectations and, therefore, we refer to them as "stakeholders" from this point on.

[3] Of course, these kinds of challenges are not unique to the use of automated scoring in the educational domain. Other complex, real-world applications involving NLP also have several stakeholders. For example, in the health care industry, NLP is being used to support clinical decision-making where stakeholders can include patients, medical professionals, caregivers, and healthcare institutions.

Standard 3.8	“When tests require the scoring of constructed responses, test developers and/or users should collect and report evidence of the validity of score interpretations for relevant subgroups in the intended population of test takers for the intended uses of the test scores.”
Standard 4.19	“When automated algorithms are to be used to score complex examinee responses, characteristics of responses at each score level should be documented along with the theoretical and empirical basis for the use of the algorithms.”
Standard 6.8	“Those responsible for test scoring should establish scoring protocols ... When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented.”
Standard 6.9	“Those responsible for test scoring should establish and document quality control processes and criteria. Adequate training should be provided. The quality of scoring should be monitored and documented. Any systematic source of scoring errors should be documented and corrected.”

Table 1: Some standards in the 2014 Standards for Educational and Psychological Testing that are relevant for automated scoring.

- **Score Users**. The people to whom any assessment is administered (test takers) are important stakeholders since any decisions made about the scoring of the assessment affect them directly. This is particularly true in cases where the assessment is likely to have a significant impact on the test-takers’ futures, e.g., by contributing to a decision about whether to admit them into a college or graduate school (IELTS, GRE, TOEFL, SAT, ACT, etc.), or a decision whether to grant them a license to engage in a professional activity such as teaching (PRAXIS). If an automated scoring system is being employed to score their responses to such critically positioned assessments, test-takers want to ensure that such a system measures the same set of characteristics and skills that a well-trained human scorer would[4] and that such measurements are calculated accurately based on their responses. In addition, the test-takers also place emphasis on receiving their scores quickly and on a *score report* providing useful information on why they received the particular scores that they did and what, if anything, they can do to improve their scores if they decide to retake the assessment. Another important consumer of test scores are institutions who use the test scores to make decisions. For example, universities may use test scores to make placement decisions; immigration authorities may use test scores to make visa decisions; school districts and states may use test scores to make funding and policy decisions. Such institutions want to ensure that the scores from the test are valid and reliable and that any automated scoring component does not introduce any biases towards any particular subgroup of test takers.

- **Teachers**. Teachers in classrooms are also likely to be affected by the decision to use automated scoring systems. It is important to them as stakeholders that scores from automated systems are not used inappropriately. For example, if the assessment — and the automated scoring system — are only designed to measure the students’ writing proficiency, the scores assigned by the system should not be used for a different purpose, e.g., to assess the teacher’s teaching abilities. Teachers are also impacted if automated systems are used directly in the classroom, e.g., to provide feedback to students in the context of*formative* assessments (informal assessments conducted by teachers in the classroom to improve how students learn). For such cases, the teachers want to ensure that the feedback provided by the system is reasonably accurate, does not lead the students astray from the actual learning goal, and encourages engagement with the material being taught.

- **Subject-matter Experts**. Subject-matter experts, also known as assessment developers, write the questions that are included in the assessments. In addition, they also assemble specific questions

[4] An automated system cannot “read” the response in the same way that a human can but it can use features that are reasonable approximations for factors that human scorers consider in their evaluation.

into assessments while taking into account that the chosen questions should cover a wide range of skills that are to be measured by the assessment (usually known as the *construct*) and that different questions try to measure complementary aspects of such skills. As part of the assessment design, they also create what is called a *scoring rubric* - a document that attempts to describe specific and consistent criteria for how human scorers should score responses to each question in the assessment. Rubrics tend to be complex and subjective, particularly for assessments that contain relatively open-ended writing or content-based questions. Such experts would like to ensure that any automated scoring system deployed to score the assessments they have designed pays attention to the scoring rubric and that only construct-relevant information is used by the system during the scoring process. For example, although the length of a response to a question designed to measure the test-taker's writing proficiency might correlate very highly with its score, it is not actually relevant to the mental construct of writing proficiently and, therefore, should not be used as a feature in the automated scoring system (Bejar, 2017).

- **Business Units**. It takes significant resources to develop, administer, and score well-designed assessments. Therefore, the institutions undertaking this process are more likely to be educational technology companies with dedicated staff for assessment development, psychometric analysis, and natural language processing. Therefore, there is a business aspect to educational assessments in addition to research. For example, business units at such companies might comprise of people who try to procure state and federal contracts for developing and scoring K-12 assessments. When it comes to automated scoring systems, such units place emphasis on more practical business aspects of system development, such as building systems that are fast (e.g., with a short turn-around time between the submission of the response and the production of the score), cost-effective to deploy and maintain in an operational setting, and have little to no measurable impact on the overall validity and reliability of the assessment.

- **NLP Researchers & Developers**. NLP researchers and developers such as the authors (and, most likely, the readers) of this paper tend to have a different perspective when it comes to building automated scoring systems. The automated scoring system should perform accurately where accuracy is generally defined as agreement with human scores. Secondly, the system should not only build on top of state-of-the-art ideas and tools from the field but, if possible, should also advance the field forward by sharing and disseminating tools, lessons, and ideas at conferences and workshops. Another important consideration is the modularity of the system that not only allows replacement of NLP components (e.g., taggers and parsers) with newer and better-performing versions but also allows new scoring features to easily be incorporated into the system. Finally, the system should be easy to maintain and well-documented to make it easy for new developers and researchers to become familiar with the system.

Since each set of stakeholders is trying to optimize the automated scoring system for a different set of criteria, it is only natural that many of the above perspectives can be at odds with each other. Below we specifically outline the conflicts between the perspectives of the **NLP researchers** and the other four sets of stakeholders.

(a) **vs. Business Units**. Business units might sometimes place a greater emphasis on getting systems to market faster and with a limited investment depending on the budget available. However, in order to build a system that has a reasonably high agreement with human scores and is more likely to generalize to new and unexpected responses in the field, NLP researchers might require additional time as well as investment. For example, annotation of additional responses may be required for use as training data, or an existing research technique from the literature may need to be adapted for the domain before deployment.

(b) **vs. Score Users**. As described above, one of the most important considerations for test-takers is a reasonably clear explanation of why they received the particular scores that they did. The NLP researchers optimizing for agreement with human scores might lean towards using more sophisticated machine learning models such as SVMs with non-linear kernels and deep neural networks. However, such models do not really lend themselves to *post-hoc interpretability* (Lipton, 2016). Although interpretability is an active area of research in the machine learning literature (Ribeiro et al., 2016; Koh and Liang, 2017; Doshi-Velez and Kim, 2017), it currently lags far behind the research on machine learning methods. Ensuring that there are no biases in automated scores – important for institutions using test scores to make decisions – is a topic that sees little discussion in the NLP literature. This is partly driven by a lack of demographic data available in publicly available datasets, as well as perhaps a focus on empirical accuracy.

(c) **vs. Subject-matter Experts**. Subject-matter experts or assessment developers want to ensure that all the hard work that has been done on their end to develop a valid (actually measuring what it is supposed to measure) and reliable assessment (scores are comparable across repeated administrations) is not undone by an automated scoring system that is either using construct-irrelevant factors as features or using a set of features that, taken together, have low *construct coverage*, i.e., they only measure part of the skill being assessed. It is difficult for NLP researchers to convert the salient aspects of a complicated – and subjective – document like the scoring rubric into features that are reasonably efficient to compute. Although the conversion can be aided by asking human scorers how they mentally translate the rubric into specific scoring decisions, humans are not as interpretable as one might think (Lipton, 2016).

(d) **vs. Teachers**. To make sure that automated scoring (or feedback) systems behave as expected if deployed for in-classroom use, NLP researchers would like to conduct research studies with such systems in real classrooms in order to collect useful data, e.g., written or spoken responses, student behavior, and indicators of engagement which can then be used to improve the system further (Burstein et al., 2016; Burstein and Sabatini, 2016; Madnani et al., 2016). However, teachers want to ensure that such systems are sufficiently nuanced — and not too primitive — to handle interactions with students and do not lead to students being distracted instead of learning. Furthermore, it takes time to build up a level of trust between the teachers and NLP researchers as a system is being fine-tuned and developed.

It is evident from the above discussion that trying to cater to everyone is akin to solving a difficult constraint satisfaction problem. For example, if NLP researchers want to build a more interpretable automated scoring system that can provide more useful feedback to test-takers, it requires investing more money and time which might need to be negotiated with the business units. Or, if subject-matter experts design an assessment with more intricate, open-ended questions in order to accurately assess the required set of skills, the corresponding automated scoring system would likely require more time and resources to build and potentially be less transparent.

5 Our Position

While NLP and machine learning techniques form a core part of automated scoring systems, there are several other non-NLP considerations that need to be taken into account when designing and developing such techniques. It is important to remember the perspectives of the other stakeholders — as outlined in the previous section — when making decisions during the NLP component development, and not simply focus on having the most accurate, or fastest automated scoring systems. While it is admirable to take advantage of the recent availability of relevant data and develop novel and more sophisticated NLP techniques to advance the field, sometimes it is necessary to take a step back and ensure that the NLP advances are also aligned with the interests of other stakeholders in automated scoring.

For example, the NLP researchers may find through ablation studies that certain computationally complex features are not contributing much to the overall accuracy of the model. They may choose to simplify the processing pipeline and resulting models by excluding those features, at very little cost to overall accuracy (on some held-out evaluation data). However, if those features are measuring important aspects of the construct, removing them weakens the validity and construct coverage of the system, since now the system omits measuring an important aspect of what is being measured by humans, as defined in the scoring rubric. This is very important for both subject-matter experts (who will assume that any automated system is at least trying to measure the same thing that humans are) and business units (who require a valid test).

6 Case Studies

To explain our position more concretely, we describe three hypothetical case studies involving automated scoring or feedback systems. We focus on some of the stakeholder interactions that we believe are critical for successful deployment of automated scoring systems and provide suggestions for best practices. Note that many of our suggestions are adapted from the best practices that are already recommended by research practitioners in the educational measurement community.

6.1 Adding Automated Scoring to an Existing Assessment

The decision to add automated scoring to an existing assessment can be initiated by the business units as a potential cost-saving measure. It can also be initiated by the NLP researchers who believe they have developed a system that can accurately score a particular type of spoken or written response. In either case, a number of considerations need to be taken into account before automated scoring can be added to an existing assessment.

If a business unit wants to add automated scoring for an existing assessment, they should typically first approach the NLP researchers to estimate the effort involved in developing automated scoring capabilities for the specific question types contained in the assessment (e.g. essays that measure writing quality or free-text responses that measure knowledge of some content area). NLP researchers will assess the feasibility of automated scoring for the requested item types. At this point, it is critical to engage the subject-matter experts in order to fully understand the construct being measured, the scoring rubric and any supplemental scoring guidelines, as well as to get access to any training materials used to train human scorers. Without engaging the subject-matter experts it is all too easy to make assumptions about the assessment based on observations made from a limited amount of scored data. This can lead to automated scoring systems that measure the construct inaccurately or in a limited fashion, that handle aberrant responses incorrectly, and that could ultimately lead to unfairly scored assessments yielding significant consequences for the test-takers.

If NLP researchers initiate the request to add automated scoring to an assessment, they should have already connected with the subject-matter experts to ensure that they have built a system that adequately measures the correct construct. It is important that they also consider other aspects of the assessment and communicate with all relevant stakeholders. For example, if an assessment has a low number of test-takers each year, then the amount of data available to monitor[5] the automated system may be too low. If the number of responses available for monitoring is low, then the risks are that a sample small enough for the monitoring to be cost effective will not provide statistically meaningful monitoring metrics. Conversely, a sample large enough to provide statistically meaningful metrics would effectively offset any potential cost-savings obtained from automated scoring. This consideration would be very important to the business units who have to fund both the setup costs of integrating automated scoring into an existing assessment, as well as the ongoing maintenance costs.

Finally, for any automated scoring system that is proposed, NLP researchers need to take into account ethical considerations regarding fairness and validity and evaluate the system on dimensions other than just the agreement with human scores. For example, it is critical to evaluate whether the system is biased towards certain sub-populations of test-takers. Aggregated metrics of agreement with human scores (such as Pearson correlation or Cohen's kappa) will not be able to capture such biases or fairness issues. Madnani et al. (2017) proposed an open-source tool, RSMTool[6], as an initial step in this direction. In

[5] Monitoring is an important aspect of high-stakes automated scoring systems, wherein a random sample of the responses scored by the automated system are also scored by an expert human scorer to ensure that the system is performing as expected.

[6] https://github.com/EducationalTestingService/rsmtool

addition to the agreement metrics with human scorers, RSMTool also evaluates how well the automated scoring system performs for each sub-population represented in the data. Of course, one key assumption underlying this type of evaluation is that the demographic information about the candidates is available and permitted to be used as part of the evaluation pipeline.

6.2 Creating a New Assessment that Includes Automated Scoring

This particular scenario offers many more opportunities for a collaborative and multi-perspective development process to be adopted right from the start as compared to the previous case study. As a business unit puts together the plans for a new assessment, automated scoring is often a desired component (usually for perceived cost-saving measures). It is important at this juncture for the subject-matter experts and the NLP researchers to collaborate in order to understand the types of questions that will be included in the new assessment, and which ones might actually be suitable for being scored automatically. It is important for the NLP researchers to understand the specific constructs being targeted by the new assessment and advise the subject-matter experts where construct-irrelevant differences might impact automated scoring.

For example, a question might use a passage about a young student with a name that is not likely to be familiar to the population of students taking the test. This means that the test-takers may guess the gender of this student, resulting in a range of pronouns appearing in the sample responses from which automated scoring models might be built. This unnecessarily “dilutes” the vocabulary of the response pool with construct-irrelevant variation. On the other hand, giving the hypothetical student a name that has a widely acknowledged gender associated with it will help limit this kind of variation without any impact on the construct coverage or validity of the question.

At the same time, while subject-matter experts are developing the questions, it is important for the NLP researchers to assess the feasibility of automated scoring for the question types being considered for automated scoring. If a question would require NLP techniques that are in very early stages of research or have not yet been fully tested, this information needs to be shared as quickly as possible with the business units so that they can build contingency plans into their budgets and timetables.

Finally, a comprehensive evaluation of the sort described in the previous case study is still warranted in this scenario since ensuring the accuracy and fairness of an automated scoring system does not depend on whether it is being deployed for a entirely new assessment or for an existing one.

6.3 Including Automated Scoring in a Classroom Setting

It is crucial that NLP researchers engage with both teachers as well as students when developing tools to be deployed in an authentic classroom setting, e.g., tools that can provide feedback on students’ writing or content knowledge. An NLP researcher designing and implementing a classroom tool in isolation is not likely to be successful, no matter the accuracy or novelty of the underlying NLP techniques. Before developing any tools, the NLP researchers need to fully understand the problem they are trying to solve by engaging teachers to find out how NLP technology can be integrated in a supportive fashion to their teaching curricula. Ideally, classroom tools should be built in an iterative fashion, by learning what features or techniques improve student engagement and learning and which ones do not. An example study that evaluates the effect of a new tool in an authentic classroom setting has many non-trivial (and non-NLP-specific) components:

- Identify a representative sample of schools/classes/students for the study
- Conduct surveys of teachers and control groups
- Conduct teacher training
- Analyze teachers’ daily/weekly logs
- Conduct classroom observations
- Conduct pre- and post-study evaluations

- Analyze tool usage logs

This requires a lot of time and effort both from the teachers as well as the NLP researchers, but at the end of the day is more likely to lead to truly useful tools that can have a positive impact on the classroom learning process and improve students' understanding of the material being taught by the teacher.

7 Conclusion

Our goal in this paper is to bring attention to the inherent complexity of automated scoring of student responses, given the large number of stakeholders, often times with different priorities. We take the position that in order to build fair and accurate automated scoring systems — especially for use in high-stakes assessments where the consequences of being unfair and inaccurate can be severe for the test-takers — NLP researchers must incorporate the perspectives of other stakeholders into the research and development process and avoid working in a bubble surrounded only by scored data and machine learning algorithms. We feel encouraged that our position is increasingly shared by many educational NLP researchers, as is evident by publications in more focused educational NLP workshops, and hope that all NLP researchers will keep this in mind as they develop new and exciting techniques for automated scoring that could be deployed for operational use. We hope to see more and more publications at NLP venues that describe these new advances while at the same time taking into account the perspectives of many of the stakeholders described in this paper.

In closing, we would like to make some concrete suggestions – informed by our position on automated scoring – that can actually apply more broadly to the field:

1. Conference and workshop organizers should strive to include industry tracks with published proceedings so that issues of the sort presented in this paper can be discussed and appropriate best practices can be developed. In addition, this would create a resource that the NLP community at large can go to when considering issues that affect the transition of NLP technology from theory into practice. More panels at conferences with both industry practitioners and academic researchers as panelists will also be helpful.

2. Shared task organizers should provide more context around the data and the specific task. As an example, for automated scoring, additional metadata can be included, if possible, to encourage more comprehensive evaluations of the sort described in §6.1 and §6.2.

3. Companies that have experience with transitioning research into practice should share more of their experiences publicly. Such companies should also offer internship positions for students to help expose them to the challenges that accompany such a transition.

4. Faculty members should encourage students to attend industry panels at conferences and to actively pursue internships in the industry in order to learn about what it takes to for academic research to be deployed in practice. This would likely lead to better dissertations and publications with improved evaluations and nuanced discussions.

Acknowledgments

We would like to thank Jill Burstein, Yoko Futagi, Beata Beigman Klebanov, and the three anonymous reviewers for useful insights and suggestions on how to improve the paper.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, et al. 2014. *Standards for Educational and Psychological Testing*. American Educational Research Association.

Solon Barocas, Sorelle Friedler, Joshua Kroll, Berk Ustun, Suresh Venkatasubramanian, and Hanna Wallach, editors. 2017. *Proceedings of the ACM SIGKDD Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Sumit Basu, Chuck Jacobs, and Lucy Vanderwende. 2013. Powergrading: a Clustering Approach to Amplify Human Effort for Short Answer Grading. *Transactions of the Association for Computational Linguistics (TACL)*, 1:391–402.

Beata Beigman Klebanov, Jill Burstein, Judith Harackiewicz, Stacy Priniski, and Matthew Mulholland. 2016. Enhancing stem motivation through personal and communal values: Nlp for assessment of utility value in student writing. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 199–205, San Diego, CA, June. Association for Computational Linguistics.

Isaac I Bejar, Robert J Mislevy, and Mo Zhang. 2016. Automated Scoring with Validity in Mind. *The Wiley Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications*, page 226.

Isaac I Bejar. 2017. Threats to Score Meaning in Automated Scoring. *Validation of Score Meaning for the Next Generation of Assessments: The Use of Response Processes*, page 75.

Jared Bernstein, A. Van Moere, and Jian Cheng. 2010. Validating Automated Speaking Tests. *Language Testing*, 27(3):355–377.

Brent Bridgeman, Catherine Trapani, and Yigal Attali. 2012. Comparison of Human and Machine Scoring of Essays: Differences by Gender, Ethnicity, and Country. *Applied Measurement in Education*, 25(1):27–40.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The Eras and Trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117.

Jill Burstein and John Sabatini. 2016. The Language Muse Activity Palette: Technology for Promoting Improved Content Comprehension for English Language Learners. In *Adaptive Educational Technologies for Literacy Instruction*, chapter 17, pages 275–280. Taylor & Francis, Routledge: NY.

Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu, Martin Chodorow, Lisa Braden-Harder, and Mary Dee Harris. 1998. Automated Scoring Using A Hybrid Feature Identification Technique. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 206–210, Montreal, Quebec, Canada, August. Association for Computational Linguistics.

Jill Burstein, Martin Chodorow, and Claudia Leacock. 2004. Automated Essay Evaluation: The Criterion Online Writing Service. *AI Magazine*, 25(3):27.

Jill Burstein, Norbert Elliot, and Hillary Molloy. 2016. Informing Automated Writing Evaluation using the Lens of Genre: Two studies. *CALICO journal*, 33(1):117.

Brian E. Clauser, Michael T. Kane, and David B. Swanson. 2002. Validity Issues for Performance-Based Tests Scored With Computer-Automated Scoring Systems. *Applied Measurement in Education*, 15(4):413–432.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based Recurrent Convolutional Neural Network for Automatic Essay Scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162, Vancouver, Canada, August. Association for Computational Linguistics.

Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *CoRR*, abs/1702.08608.

Myroslava O. Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge. **SEM 2013: The First Joint Conference on Lexical and Computational Semantics*.

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. 2016. Coarse-grained Argumentation Features for Scoring Persuasive Essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 549–554, Berlin, Germany, August. Association for Computational Linguistics.

Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894, International Convention Centre, Sydney, Australia, 06–11 Aug. PMLR.

Zachary C. Lipton. 2016. The Mythos of Model Interpretability. In *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning*.

Luca Lugini and Diane Litman. 2017. Predicting specificity in classroom discussion. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–61, Copenhagen, Denmark, September. Association for Computational Linguistics.

Nitin Madnani, Jill Burstein, John Sabatini, Kietha Biggers, and Slava Andreyev. 2016. Language muse: Automated linguistic activity generation for english language learners. In *Proceedings of ACL-2016 System Demonstrations*, pages 79–84, Berlin, Germany, August. Association for Computational Linguistics.

Nitin Madnani, Anastassia Loukina, Alina von Davier, Jill Burstein, and Aoife Cahill. 2017. Building Better Open-Source Tools to Support Fairness in Automated Scoring. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 41–52, Valencia, Spain, April. Association for Computational Linguistics.

Courtney Napoles and Chris Callison-Burch. 2015. Automatically scoring freshman writing: A preliminary investigation. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 254–263, Denver, Colorado, June. Association for Computational Linguistics.

Ulrike Pado. 2017. Question difficulty – how to estimate without norming, how to use for automated grading. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–10, Copenhagen, Denmark, September. Association for Computational Linguistics.

Ellis B. Page. 1966. The Imminence of ... Grading Essays by Computer. *The Phi Delta Kappan*, 47(5):238–243.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chong Min Lee. 2017. Investigating Neural Architectures for Short Answer Scoring. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 159–168, Copenhagen, Denmark, September. Association for Computational Linguistics.

Swapna Somasundaran, Chong Min Lee, Martin Chodorow, and Xinhao Wang. 2015. Automated Scoring of Picture-based Story Narration. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 42–48, Denver, Colorado, June. Association for Computational Linguistics.

Md Arafat Sultan, Cristobal Salazar, and Tamara Sumner. 2016. Fast and Easy Short Answer Grading with High Accuracy. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1070–1075, San Diego, California, June. Association for Computational Linguistics.

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*, pages 1882–1891, Austin, Texas, November. Association for Computational Linguistics.

Yi Tay, Minh C. Phan, Luu Anh Tuan, and Siu Cheung Hui. 2017. SkipFlow: Incorporating Neural Coherence Features for End-to-End Automatic Text Scoring. *CoRR*, abs/1711.04981.

David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. A Framework for Evaluation and Use of Automated Scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13.

Joshua Wilson and Trish Martin. 2015. Using pegwriting® to support the writing motivation and writing quality of eighth-grade students: A quasi-experimental study. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 179–189, Denver, Colorado, June. Association for Computational Linguistics.

Victoria Yaneva, Constantin Orasan, Richard Evans, and Omid Rohanian. 2017. Combining multiple corpora for readability assessment for people with cognitive disabilities. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 121–132, Copenhagen, Denmark, September. Association for Computational Linguistics.

Yongwei Yang, Chad W. Buckendahl, Piotr J. Juszkiewicz, and Dennison S. Bhola. 2002. A Review of Strategies for Validating Computer-Automated Scoring. *Applied Measurement in Education*, 15(4):391–412, oct.

Nasser Zalmout, Hind Saddiki, and Nizar Habash. 2016. Analysis of foreign language teaching methods: An automatic readability approach. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, pages 122–130, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M. Williamson. 2009. Automatic Scoring of Non-native Spontaneous Speech in Tests of Spoken English. *Speech Communication*, 51(10):883–895.

Torsten Zesch, Michael Wojatzki, and Dirk Scholten-Akoun. 2015. Task-Independent Features for Automated Essay Grading. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 224–232, Denver, Colorado, June. Association for Computational Linguistics.