Building Robust Content Scoring Models for Student Explanations of Social Justice Science Issues

Allison Bradford and Marcia C. Linn
Berkeley School of Education
University of California, Berkeley
{allison_bradford, mclinn}@berkeley.edu

Kenneth Steimel and Brian Riordan
ETS
{ksteimel, briordan}@ets.org

Abstract

With increased attention to connecting science topics to real-world contexts, like issues of social justice, teachers need support to assess student progress in explaining such issues. In this work, we explore the robustness of NLP-based automatic content scoring models that provide insight into student ability to integrate their science and social justice ideas in two different environmental science contexts. We leverage encoder-only transformer models to capture the degree to which students explain a science phenomenon, understand the intersecting justice issues, and integrate their understanding of science and social justice. We developed models training on data from each of the contexts as well as from a combined dataset. We found that the models developed in one context generate educationally useful scores in the other context. The model trained on the combined dataset performed as well as or better than the models trained on separate datasets in most cases. Comparing human scores with the automated scores using quadratic weighted kappas demonstrate that these models perform above the threshold for use in classrooms.

1 Introduction

This study investigates the robustness of Natural Language Processing (NLP)-based automatic content scoring models that assess secondary school students' ability to integrate their science and social justice ideas to explain social justice science issues (SJSI; Morales-Doyle, 2017) in two different contexts. In particular, we investigate the robustness of content scoring models in terms of their ability to score out-of-distribution responses as we attempt to generalize the models from one SJSI context to another. The contexts are (a) a unit about combustion reactions and asthma caused by exposure to particulate matter pollution and (b) a unit about global climate change and exposure to extreme heat occurring in urban heat islands. In both units, which are also aligned to state science standards, the students explore the racially disparate impacts of the environmental hazard (particulate matter pollution, extreme heat in urban spaces). In the units, students are supported to explore typical disciplinary content and make connections to local justice issues. They answer the *Impacts Item*, explaining whether all people are impacted by the environmental hazard in the same way. We explore the possibility of building a robust domain general model that can be used across multiple SJSI contexts.

The curriculum, assessments, and scoring rubrics were developed by a research practice partnership (RPP) including classroom teachers, computer scientists, and learning scientists guided by the Knowledge Integration pedagogy (KI; Linn and Eylon, 2011). The automatic content scoring models were created to assess the degree to which students connect their understanding of the environmental concepts with understanding of the social justice issues when explaining whether everyone is impacted in the same way. As teachers reformulate their instruction to include social justice perspectives, automatic content scoring models can help teachers by capturing student progress. They are especially valuable for social justice ideas that might be new to science teachers. We investigate the accuracy and robustness of automatic content scoring models that can quickly assess student explanations, particularly when those explanations contain social justice ideas. In this study we ask:

- Can we develop NLP models that accurately capture students' integrated understanding of SJSIs, as measured by human-computer agreement?
- What are the affordances and limitations of combining training datasets from different disciplinary contexts to develop robust automatic content scoring models of SJSI?

2 Related Work

This study builds on prior research integrating social justice into science curriculum and leveraging AI techniques to score student essays.

2.1 Social Justice Science Issues (SJSI)

In this study, we combined social justice science pedagogy (Morales-Doyle, 2017) with Knowledge Integration (KI) design framework (Linn and Eylon, 2011) to design units featuring SJSI. Centering issues of social justice in science teaching and learning offers promise for preparing students to deal with contemporary science issues. One productive example involves grounding science teaching in local social justice science issues (Morales-Doyle, 2017). Students in this Chicago neighborhood drew attention to the contamination in the soil in the community garden. Introducing SJSIs provides opportunities for students to make sense of issues impacting their own communities and raises issues around inequality and racism (Morales-Doyle et al., 2019). In making sense of such issues, students connect typical science ideas to interpret how an environmental phenomenon impacts their community. This enables them to integrate disciplinary ideas with social justice ideas to explain why the impacts are different across racial and socioeconomic groups.

We also developed an aligned assessment, the *Impacts* item, that requires students to explain an environmental hazard and whether all people are impacted by it the same way. For example, when explaining who is impacted by urban heat islands, a student wrote, *"I don't think all people are impacted by the effects of climate change in the same way. Red areas on the map are 5-20 degrees higher than blue or green areas. Red areas are mostly habited by brown and black people. Red areas have less funding because of segregation and racism. They have less access to government funds and less green areas which help with the decrease in climate change."* To assess student explanations, the KI framework indicates that assessment should focus on the integration of concepts rather than the accuracy of isolated ideas, requiring the development of automatic content scoring models that capture the degree to which students integrate their ideas. As such, we developed an overall KI score rubric (Table 1; Liu et al., 2008; Liu et al., 2016) as well as KI-aligned Disciplinary and Justice subscore rubrics to score training data.

2.2 Automatic Content Scoring

Automatic content scoring can be traced back to early work on the Project Essay Grader (PEG) system which leveraged computers to grade essays and found that a computer rater's score was nearly as highly correlated with human raters' scores as the human raters' scores were with each other (Page, 1966). This work paved the way for Automatic Essay Scoring (AES) models and automatic content scoring. Many advances in AES modeling have resulted in widely used classroom and high stakes assessments. For example the e-rater automated scoring system is used for the Graduate Management Admission Test (GMAT; Burstein, 2003). To score short, student-generated free-text responses such as the *Impacts item* according to a scoring rubric, c-rater has shown promise (Leacock and Chodorow, 2003). C-rater works by determining whether a natural language response is part of the set of correct ideas that could be expressed in response to the prompt. To do so, the model uses a number of natural language processing techniques to normalize a response by attending to sources of variation in expression of the same idea: syntactic variation, morphological variation, pronoun reference, the use of synonyms or similar words, and spelling or grammatical errors.

Recently, researchers working on automatic content scoring for short answer responses have sought to incorporate approaches that have been effective in the realm of AES (e.g. Riordan et al., 2017) like the use of neural architectures (e.g. Zhao et al., 2017) including pre-trained transformer models (e.g. Yang et al., 2020). In particular, we build on the automatic content scoring work of Riordan et al. (2020) which showed that recurrent neural network and encoder-only transformer models performed just as well or better than feature-based models. Riordan et al. (2020) also demonstrated that the encoder-only transformer-based models were more robust to spurious, dataset-specific learning cues when applying scoring rubrics. Thus, we adopt a similar approach of fine-tuning encoder-only transformer models, BERT and SciBERT, to develop short answer scoring models for KI, Disciplinary, and Justice scores.

3 Data and Experimental Design

We developed automatic content scoring models to automatically score the *Impacts* item which is found in several units: Global Climate Change and

KI Score	Criteria	Asthma Example	UHI Example
1	Irrelevant	idk	asifhsdif
2	Vague	Yes, climate change will effect everyone in the whole world.	I think in some ways yes and in some ways no.
3	Partial link: one target idea	Yes, because if you have more freeways or factories where you live you could have more of the effects of incomplete combustion.	No, because there is less greenery, and plants and trees help to keep things cool in urban heat islands.
4	Full link: links two target ideas	People who are lower income are impacted by climate change more than people who aren't because they sometimes have to live closer to factories and other places where there could be harm.	No, some people who for example live in poorer or redlined areas will be more impacted. As those areas don't have as much greenery or architecture that can help with the heat.
5	Full links: links three or more target ideas	NO! Racially oppresed groups are affect more by climate change. These groups are in redlined communities which put near industrial areas which produce green house gases. These greenhouse gas emmisions give you a higher chance to have asthma.	Black and hispanic people who live in poorer residences have less trees and grass nearby, as an effect of redlining, which makes poorer neighborhoods hotter. The rich white neighborhoods are invested in by banks, and have much more trees and grass, making their neighborhood 5-20 degrees cooler.

Table 1: Rubric for KI score with examples from both unit contexts.

Urban Heat Islands (UHI; 9th grade) and Chemical Reactions and Asthma (Asthma; 7th grade). In this section, we describe the item, scoring rubrics, training data, and experimental design. The section that follows details our model development approach.

3.1 Assessment Item and Scoring Rubrics

The *Impacts* item asks students to explain whether all people are impacted by an environmental hazard in the same way. In both unit contexts, students connect their science understanding to the role of race, socioeconomic status, and policies like redlining in their local communities. In the UHI unit, the item prompt elicits ideas about how the Sun transfers energy to different surfaces and how those surfaces contribute to the surrounding temperature. In the Asthma context, the item prompt elicits ideas about how the products of incomplete combustion reactions relate to asthma.

To develop the scoring model for the Impacts item, we first developed a knowledge integration (KI score) rubric (scale 1-5; Liu et al., 2008; Liu et al., 2016) and two subscore rubrics: Disciplinary and Justice (scale 0-2). The KI score measures the overall integration of ideas in the student explanation and is agnostic to the explanation context (see rubric in Table 1). The Disciplinary subscore characterizes how students integrate domain specific target ideas in their explanations. While the rubric structure is the same, the disciplinary target ideas are different in the Asthma and UHI contexts. The Justice subscore characterizes how students integrate target ideas about historical policies and social injustices into their explanations. The justice target ideas are the same in the Asthma and UHI contexts. Target ideas were identified in collaboration with all members of the RPP. A subscore of 0 indicates no mention of target ideas, a subscore of 1 indicates an isolated target idea, and a subscore of 2 indicates the integration of two or more target ideas. All rubrics reward students for linking their ideas and connecting evidence, and do not penalize students for incorrect ideas.

3.2 Training Dataset and Experimental Design

We applied the scoring rubrics to data from previous classroom studies where students responded

Disciplinary Subcore	Criteria	Asthma Example	UHI Example
0	No mention	I think so	Everyone is affected
1	Isolated	Yes, because historical practices like redlining made certain neighborhoods that had poorer air quality be the only neighborhoods available to people of color.	no, some people they are homeless and have it harder when there is no shelter and it's really hot outside. when other people can go inside too and air-conditioned houses.
2	Full link	Many places are redlined and those neighborhoods are usually near freeways and refineries and have poor living conditions. People of color are often the ones forced to live in redlined areas so they deal with the incomplete combustion from the freeways and refineries much more than people who live in an area that is not redlined.	People of color and people in lower-income households are much more likely to experience the effects of a global rising temperature. They are less likely to be able to afford proper air conditioning and to live near green areas, which causes an increased rate of heat-related hospital visits and deaths.

Table 2: Rubric for Justice subscore with examples from both unit contexts.

Disciplinary Subcore	Criteria	Asthma Example	UHI Example
0	No mention	Probably	Yes because everyone lives in the world and global warming affects all parts of the planet.
1	Isolated	Depending on how many Carbon Monoxide and Particulates there are, which is influenced by factories. If you live closer or work in factories, the effect will be much worse	No, because there is less greenery, and plants and trees help to keep things cool in urban heat islands.
2	Full link	Some places have more incomplete combustion, that can make soot and carbon monoxide. This can affect the air quality that people breath in, which causes more cases of asthma or other medical conditions.	Lower-income families and neighborhoods are affected by the lack of trees and greenery to cool down the temperatures. It can affect the residents towards more respiratory diseases, heart problems, or dehydration.

Table 3: Rubric for Disciplinary subscore with examples from both unit contexts.

to the Impacts item. Available data included 1690 responses from the Asthma unit and 548 responses from the UHI unit. The student responses are short essays, typically ranging from 1-3 sentences long. The students represented in the training data are from the 6th-9th grade in schools in a large, Western United States metropolitan area.

To assess reliability of human scoring before building the models, two raters independently applied the rubrics to 5 percent of the data and then calculated Pearson's kappa to measure our interrater reliability. We discussed disagreements and refined the rubrics. We repeated the process until we achieved a kappa > 0.85 for the KI, Disciplinary, and Justice scores. The remaining data was split 50-50 among the two raters and hand scored.

Given the considerably smaller number of responses from the UHI unit, we wondered if data from the Asthma unit context could be used to supplement the data from the UHI unit context to enhance the likelihood of developing a scoring model that performs well (as measured by alignment to human scoring). With this in mind, we established three training datasets: 1) the 548 responses collected in the UHI unit context, 2) the 1690 responses collected in the Asthma unit context, and 3) the combined 2238 responses collected across both unit contexts. Descriptive statistics for KI, Disciplinary, and Justice scores for each of the training datasets can be found in Table 4. To evaluate the effect of the composition of the training dataset, we developed the three scoring models using each training dataset. This resulted in nine total models:

- UHI-trained KI
- UHI-trained Disciplinary
- UHI-trained Justice
- Asthma-trained KI
- Asthma-trained Disciplinary
- Asthma-trained Justice
- Combined-trained KI
- Combined-trained Disciplinary, and
- Combined-trained Justice.

4 Models

4.1 Modeling Approach

The human-scored data in three training datasets were used to train content scoring models for KI, Disciplinary, and Justice scores. The models were based on encoder-only transformer models (in this case, BERT and SciBERT), following prior work (Riordan et al., 2020). The models for KI, Disciplinary, and Justice scores were trained independently, with each score representing the degree of integration for the corresponding aspect of the content of the response. Models were trained on ordinal scores (1-5 for KI, 0-2 for Justice and Disciplinary) using the text in each response. The modeling approach was a standard "instance-based" approach (as opposed to similarity-based approach; c.f. Horbach and Zesch, 2019). While instance-based models may not generalize well across prompts (Horbach and Zesch, 2019), we anticipated that responses generated by UHI and Asthma versions of the *Impacts* item would succeed because many ideas or phrases associated with high level scores are the same in both unit contexts. Ideas that are specific to a particular unit context are unlikely to occur in the other context, minimizing the likelihood that words or phrases associated with a high score from one unit context would be associated with a low score in the other unit context.

We used BERT (Devlin et al., 2019) for the KI and Disciplinary scores and SciBERT for the Justice score (Beltagy et al., 2019). The backbone selection was based upon prior experimentation not reported in this paper. Following standard practice, for all models, during training, a special classification token '[CLS]' was added to the beginning of each input sequence. To make score predictions, the learned representation for the [CLS] token was processed by an additional layer with sigmoid activation, outputting a real-valued score prediction. This real value was mapped back to ordinal scores for making predictions.

During training, learning rates were tuned individually for each model using grid search. Hyperparameter optimization was carried out as follows: We trained using 10-fold cross-validation with an 80-10-10 training/validation/test split. We tuned hyperparameters by training on each train split and evaluating on validation splits. We retained the epoch where best performance was observed and the predictions from that epoch. Then, to select

Training Dataset	Mean	Median	Min	Max	Std Dev
UHI-KI	2.73	3	1	5	0.82
Asthma-KI	2.75	3	1	5	0.78
Combined-KI	2.75	3	1	5	0.79
UHI-Disciplinary	0.79	1	0	2	0.58
Asthma-Disciplinary	0.82	1	0	2	0.61
Combined-Disciplinary	0.81	1	0	2	0.60
UHI-Justice	0.21	0	0	2	0.44
Asthma-Justice	0.17	0	0	2	0.40
Combined-Justice	0.18	0	0	2	0.41

Table 4: Descriptive statistics for KI, Disciplinary, and Justice Scores for each training dataset

the best hyperparameters, we evaluated the performance of the pooled predictions across all folds of the validation sets. We trained final models by training on the combined train and validation sets, using 10-fold cross-validation and using the best-performing hyperparameters from the prior hyperparameter optimization.

4.2 Classroom Testing and Model Evaluation

After developing the models, we performed additional evaluation using a sample from newly collected classroom data. To evaluate the models on new data, we embedded the *Impacts* item at three time points in both the UHI and Asthma units: on a pretest, within the lesson about the SJSI, and on a posttest. Two ninth grade science teachers taught the UHI unit (student N= 95) and one seventh grade science teacher taught Asthma (student N = 56). We selected a balanced sample of 100 responses from each unit to evaluate the models we built. The responses were human scored and scored by each of the models. We used QWK, a measure of agreement for ordinal ratings that ranges from 0 to 1 and accounts for chance agreement (Fleiss and Cohen, 1973), to compare the performance of the scoring models trained on each training dataset.

5 Results and Discussion

5.1 RQ1: Developing a model to capture students' integrated understanding of SJSI

After model development, we evaluated each model (UHI-trained, Asthma-trained, and Combined-trained KI, Justice and Disciplinary scoring models) on 100 student responses from both the Asthma and the UHI units. The test data were collected during classroom testing and not present in the training dataset during model development. The responses were hand scored by the first author and scored by each of the models. We used quadratic weighted kappa as a metric to evaluate model performance (Table 5).

All models developed performed sufficiently well ($QWK \geq 0.70$, rounded normally; Williamson et al., 2012) in the evaluation context that corresponded to the training context, i.e Asthma-trained KI, Disciplinary, and Justice models performed sufficiently well on new data collected from student learning from the Asthma unit. UHI-trained models performed sufficiently well on new data from students learning the UHI unit. The Combined-trained models also performed sufficiently well ($QWK \geq 0.70$, rounded normally; Williamson et al., 2012) for new data collected in both the Asthma and UHI units. These results suggest that we can automatically assess student progress in explaining SJSI.

5.2 RQ2. Affordances and limitations of combining datasets to develop AES models for similar instructional contexts

In most cases, the model built on a larger training dataset performs better, even if the training dataset includes data from a different instructional context. For example, the Combined-trained Disciplinary model performed best for data from both the Asthma ($QWK = 0.9380$) and the UHI units ($QWK = 0.8273$). Additionally, the Asthma-trained models perform better or as well as UHI-trained models for test data from the UHI context. Figures 1 and 2 illustrate the trend that as more data is added to the training dataset, the QWK either remains approximately the same or increases.

An exception to this trend are the models for the Justice score (Figure 3). The Asthma-trained Justice model performs best for data from both

Training Context	Evaluation Context	KI QWK	Disciplinary QWK	Justice QWK
Asthma (N=1690)	Asthma	0.9649	0.9265	0.9323
UHI (N=548)	UHI	0.9071	0.7531	0.6983
Asthma (N=1690)	UHI	0.9137	0.7499	0.8344
UHI (N=548)	Asthma	0.7941	0.5479	0.8177
Combined (N=2238)	Asthma	0.9385	0.9380	0.8785
Combined (N=2238)	UHI	0.9432	0.8273	0.7922

Table 5: Model evaluation results (quadratic weighted kappa, QWK) on the 100 newly collected student responses for models trained on data from the Asthma context, the UHI context, and the Combined dataset.

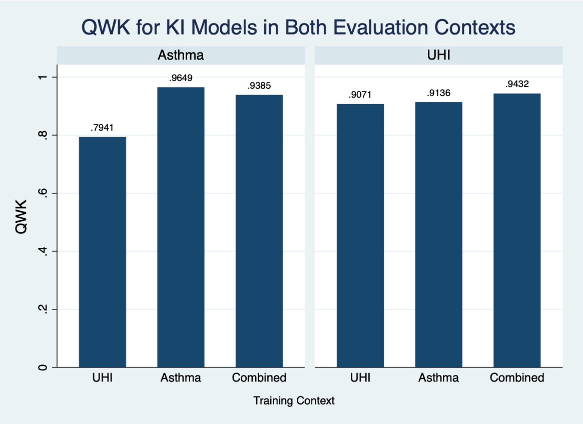

Figure 1: QWK for KI score for each model in both evaluation contexts

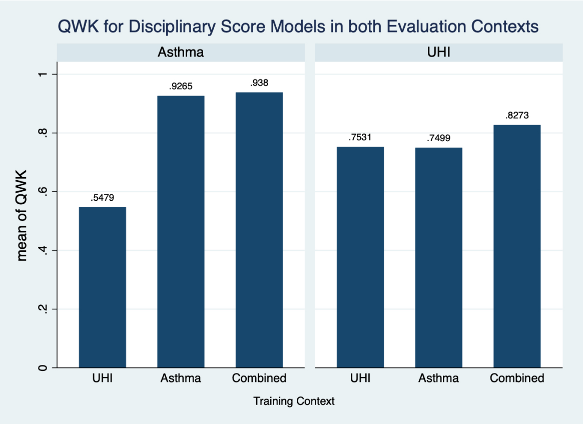

Figure 2: QWK for Disciplinary score for each model in both evaluation contexts

the Asthma unit and the UHI unit compared to the Combined-trained Justice model, even though it was trained using a smaller dataset and does not contain responses from the UHI unit. Of the 100 UHI test responses, there were six responses where the Asthma-trained Justice model accurately scored the response and the Combined-trained Justice model did not accurately score the response. In each of these responses, the Combined-trained model scored the response lower than the human rater. Four of these six responses were scored at a level 2, the highest score, by the human rater and Asthma-trained model and at a level 1 by the Combined-trained model. For example, the student explanation, “No, people are affected differently by climate change. The reasons behind it are also racially driven, as those who are affected more are likely to be people of color due to redlining and the zoning of housing” was accurately given a Justice score of 2 by the Asthma-trained Justice model and given a score of 1 by the Combined-trained Justice model. The ideas about people of color being more impacted due to historical redlining and housing policies contained in this responses are well represented in the Asthma training dataset.

With this in mind, a possible explanation for the difference in performance is that the Asthma dataset has more responses and more consistent representation of the target justice ideas. As such, it might be reasonable to expect it to perform best. Further, the justice context requires real world knowledge which is a difficult task for transformer models. Additionally, the average Justice score across the 100 UHI test responses was 0.66, while the average of the Justice scores predicted by the Asthma-trained models was 0.61, the average of the Justice scores predicted by the UHI-trained models was 0.45 and the the average of the Justice score predicted by the Combined-trained models was 0.52. The lower average predicted scores

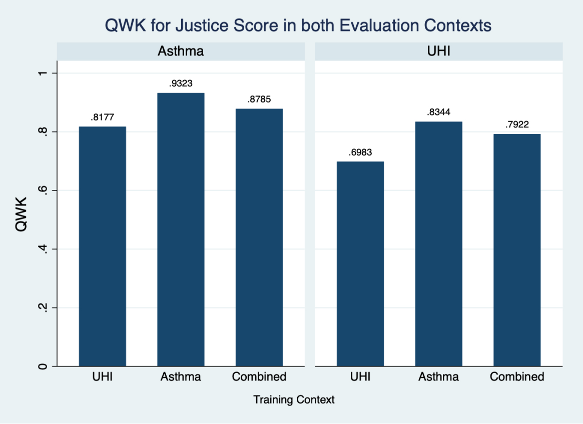

Figure 3: QWK for Justice score for each model in both evaluation contexts

from the UHI-trained and Combined-trained models might indicate that the justice ideas represented in the UHI training dataset are not well-aligned to the justice ideas expressed in the newly collected UHI test set.

Despite some reductions in performance, the Combined-trained KI, Justice, and Disciplinary models all perform well enough to be used in classrooms (Williamson et al., 2012). For the UHI instructional context, where training data was limited during model development, the combined model enhances performance suggesting the promise of the modeling approach for developing a model for in multiple instructional contexts.

6 Conclusions and Next Steps

This study investigates the robustness of pedagogically aligned automatic content scoring models trained for one SJSI context when used for a different SJSI context and of the model trained on multiple SJSI. We found that the models are robust across these contexts. Models developed in one context generate educationally useful scores in the other context. The model trained on the combined dataset is as good or better than the models trained on separate datasets in most cases. These findings underscore the value of using classroom data to fine-tune encoder-only transformer models using a pedagogically-grounded scoring rubric. In particular, the models were robust for scoring student responses for knowledge integration. They also demonstrate the potential for using "instance-based" models across contexts when it is unlikely that words or phrases associated with a high score from one context would be associated with a low score in another context.

Results demonstrate that these models are above threshold for use in classrooms to give students adaptive, personalized guidance based on their essay scores. They can also be used to synthesize classroom data for teachers in real time. Thus, the automatic content scoring generates KI scores, Disciplinary scores, and Justice scores that can be displayed in class-level histograms along with illustrative student responses to help teachers monitor class progress. These results suggest promise for generalizing models across similar contexts, increasing the efficiency of design of automatic content scoring models for adaptive instructional materials.

Next steps include validating the educational value of the models in classroom settings. We plan to engage the RPP in designing and testing adaptive guidance informed by KI pedagogy for each of the automatically generated scores. In addition classroom observations and interviews with teachers are needed to understand how the scores generated by the models align with teachers' assessment of student explanations of SJSIs and how access to student scores from the models shapes their instruction.

7 Limitations

The findings of the work are limited by the nature of our experimental approach. We tested models based on the data available as opposed to systematically testing training dataset size. Further, across all training data sets, the data are imbalanced with an over representation of low Justice scores. These limitations are common constraints when working with data generated in real K-12 classroom contexts.

8 Acknowledgements

We thank Dr. Zachary Pardos for his constructive feedback on an earlier version of this work and the WISE-TELS research group for their feedback. Kenneth Steimel can now be reached at ksteimel@outlook.com and Brian Riordan can be reached at bwriordan@gmail.com.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3615–3620.

Jill Burstein. 2003. The e-rater® scoring engine: Automated essay scoring with natural language processing.

Jacob Devlin, Kenton Chang, Ming-Wei Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Andrea Horbach and Torsten Zesch. 2019. The influence of variance in learner answers on automatic content scoring. In *Frontiers in education*, volume 4, page 28. Frontiers Media SA.

Claudia Leacock and Martin Chodorow. 2003. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37:389–405.

Marcia C Linn and Bat-Sheva Eylon. 2011. *Science learning and instruction: Taking advantage of technology to promote knowledge integration*. Routledge.

Ou Lydia Liu, Hee-Sun Lee, Carolyn Hofstetter, and Marcia C Linn. 2008. Assessing knowledge integration in science: Construct, measures, and evidence. *Educational Assessment*, 13(1):33–55.

Ou Lydia Liu, Joseph A Rios, Michael Heilman, Libby Gerard, and Marcia C Linn. 2016. Validation of automated scoring of science assessments. *Journal of Research in Science Teaching*, 53(2):215–233.

Daniel Morales-Doyle. 2017. Justice-centered science pedagogy: A catalyst for academic achievement and social transformation. *Science Education*, 101(6):1034–1060.

Daniel Morales-Doyle, Tiffany Childress Price, and Mindy J Chappell. 2019. Chemicals are contaminants too: Teaching appreciation and critique of science in the era of next generation science standards (ngss). *Science Education*, 103(6):1347–1366.

Ellis B Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Brian Riordan, Sarah Bichler, Allison Bradford, Jennifer King Chen, Korah Wiley, Libby Gerard, and Marcia C Linn. 2020. An empirical investigation of neural methods for content scoring of science explanations. In *Proceedings of the fifteenth workshop on innovative use of NLP for building educational applications*, pages 135–144.

Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chungmin 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.

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.

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569.

Siyuan Zhao, Yaqiong Zhang, Xiaolu Xiong, Anthony Botelho, and Neil Heffernan. 2017. A memory-augmented neural model for automated grading. In *Proceedings of the fourth (2017) ACM conference on learning@ scale*, pages 189–192.