Automatically Scoring Tests of Proficiency in Music Instruction

Nitin Madnani **Aoife Cahill** **Brian Riordan**

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

Abstract

We present preliminary work on automatically scoring constructed responses elicited as part of a certification test designed to measure the effectiveness of the test-taker as a K-12 music teacher. This content scoring differs from most previous work in that the responses are relatively long and are written by an adult population of generally proficient English writers. We obtain reasonably good scoring performance for all the test questions using simple features. We carry out some initial error analysis and show that there is still room for improvement.

1 Introduction

In this paper, we examine the feasibility of automatically scoring content-based questions from a teacher certification test which measures the test-taker's effectiveness as a K-12 music teacher. The test was designed by experts with extensive experience in music education, who consult regularly with music teachers and music education professors throughout the USA to ensure the appropriateness and validity of individual test questions.

Specifically, this test measures indicators of the beginning educator's professional readiness to teach K-12 music in each of the three major music education specialties: general, instrumental, and vocal music education. The typical test population consists of undergraduates who have completed, or nearly completed, a music education program. Materials appearing on the test reflect instrumental, vocal, jazz, and general music instruction specialties across the K-12 grade range. Note that the test contains a combination of multiple-choice as well as constructed response (essay-style) questions. A final score for the test is computed by combining the scores for all questions and *only* that score is reported to the test-takers, not individual question scores. In this paper, we focus on building automated scoring models for the essay-style questions.

2 Data

We obtained test-taker responses written between 2013 and 2015 to multiple administrations of the test. Each student wrote answers to three essay-style questions. We look at a total of six different essay-style questions across all test forms. Although we cannot disclose the actual questions for reasons of test security, Figure 1 shows a sample question from the test. It asks the prospective teacher to examine a given vocal music sample and answer questions relevant to teaching the sample to a hypothetical class of students. Overall scores are assigned on a 0–3 scale, based on the degree to which the test-taker accurately responds to the three subparts of the question. Figure 3a shows the total number of scored responses available (N) for each of the six questions.

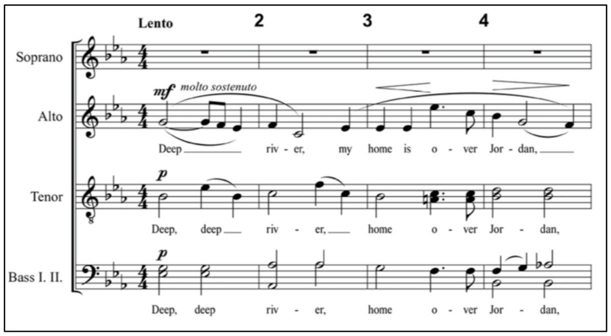
The responses to all six questions on the test are scored by two human experts on a 0–3 scale.[1] Figure 2 shows the distributions of the response lengths and the scores assigned to the responses by the first human expert (hereafter referred to as the H1 score).

3 Related Work

The test we examine here has been designed primarily to elicit content knowledge from prospective teachers in the context of instruction. Our work can be consid-

[1] For each response, the two experts are chosen randomly from a pool of 9 experts.

Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pages 217–222,
San Diego, California, June 16, 2016. ©2016 Association for Computational Linguistics


Lento
2
3
4
Soprano
Alto
Tenor
Bass I. II.
mf molto sostenuto
p
p
Deep riv - er, my home is o - ver Jor - dan,
Deep, deep riv - er, home o - ver Jor - dan,
Deep, deep riv - er, home o - ver Jor - dan,

Vocal Music (Suggested time—10 minutes)

• Examine the music example shown. Based on the technical demands of the example, choose the ensemble for which the example is most appropriate:

(a) Middle school mixed chorus
(b) High school mixed chorus

Indicate your skill-level choice at the beginning of your response.

• Identify and describe a significant performance challenge in the example. Specify the location of the challenge using measure numbers.

• Describe an appropriate rehearsal technique to assist students in meeting the challenge.

Figure 1: A sample question from the music teaching proficiency test. Note that only a part of the entire music sample included with the question is shown here.

ered similar in spirit to some of the previous work on short-answer scoring, where the focus is on scoring content-driven responses to math, biology, or computer science questions (Sukkarieh and Stoyanchev, 2009; Sukkarieh et al., 2011; Mohler et al., 2011; Dzikovska et al., 2013; Ramachandran et al., 2015; Sakaguchi et al., 2015; Zhu et al., 2016).[2] However, we claim that there is very little in that body of previous work that focuses on responses exhibiting *all* of the following characteristics:

- The responses we examine are written by a population of adults that are generally proficient English writers. This is different from most previous work where the population is generally composed of middle- and high-school students with varying levels of English proficiency.

- The average length of these responses is approximately 220 words which is much longer than the responses considered in much of the previous work (10-100 words long, on average).

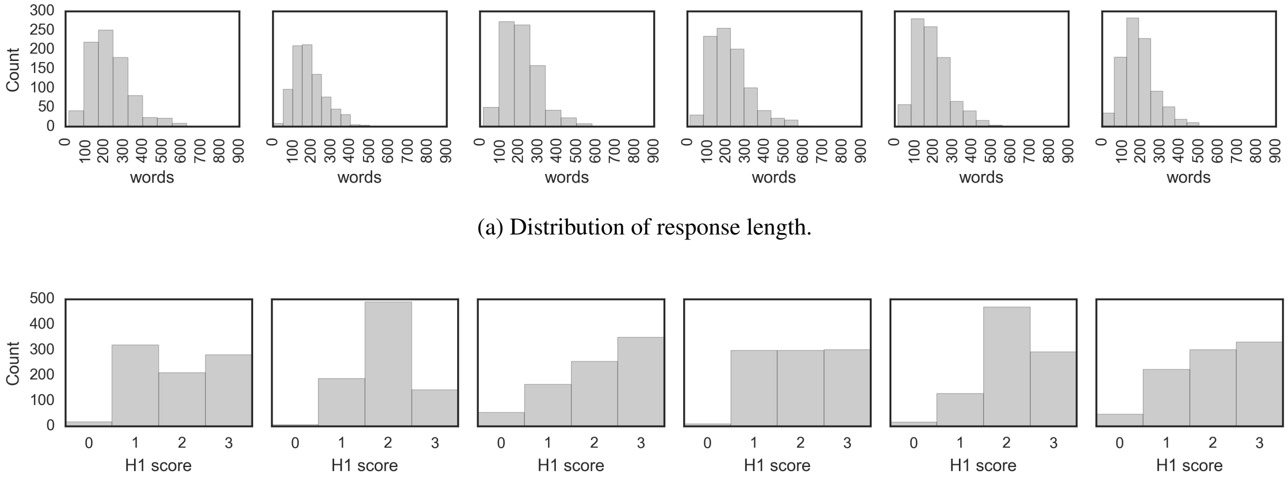
- These responses are from a high-stakes test (teacher certification) whereas previous work has focused mostly on responses from low- to medium-stakes tests (in-class discussions, homework assignments, placement tests, etc.).

The work that could be considered most similar to ours is that of Alfonseca and Pérez (2004) (further described by Pérez-Marín and Pascual-Nieto (2011)) which focuses on scoring responses to computer science questions (50-130 words long, on average) written by undergraduate students. However:

1. Their responses were written to tests that lacked the instructional context — tests that assessed content comprehension but *not* how that content can best be taught to a class of K-12 students. Although both types of responses might require understanding the same concepts, they are likely to be expressed differently. For example, the following is an excerpt from a sample response to the question in Figure 1.

   "This example is best suited for a high school mixed chorus. One performance challenge that would be likely for a HS chorus performing this work would be the octave leap in the alto part in measure 3. (This is also found in measure 11.) This passage needs to maintain the legato phrasing marked throughout and needs to crescendo smoothly without a loss of tone and without accenting the top E-flat. Students may tend to restrict their throats in order to

[2] See Table 3 in Burrows et al. (2015) for a comprehensive list.



(a) Distribution of response length.

(b) Distribution of the response scores assigned by the first human expert (H1 scores).

Figure 2: The response length and H1 score distributions of all six questions from the test.

reach the high note. With insufficient breath support, the crescendo and legato phrasing will not be musical. . . . "

2. Their responses were scored by comparing to human-authored reference answers whereas our scoring approach does not require any reference answers.

To the best of our knowledge, there has not been any work that uses a machine-learned model to automatically score questions that measure content-based teaching ability.

4 Content Scoring Model

We split the data available for each of the questions into training and test sets with 70% for training and 30% for test. We then build an automated scoring model for each question separately using the H1 score as our target. Each scoring model uses support vector regression (Smola and Schölkopf, 2004) to estimate a function that predicts human scores from vectors of binary linguistic features. We use the implementation from the *scikit-learn* package (Pedregosa et al., 2011), with default parameters except for the complexity parameter, which is tuned using cross-validation on the data provided for training.

As features in the model, we start with the set of features that have generally been used for scoring content-based short answers in the literature:

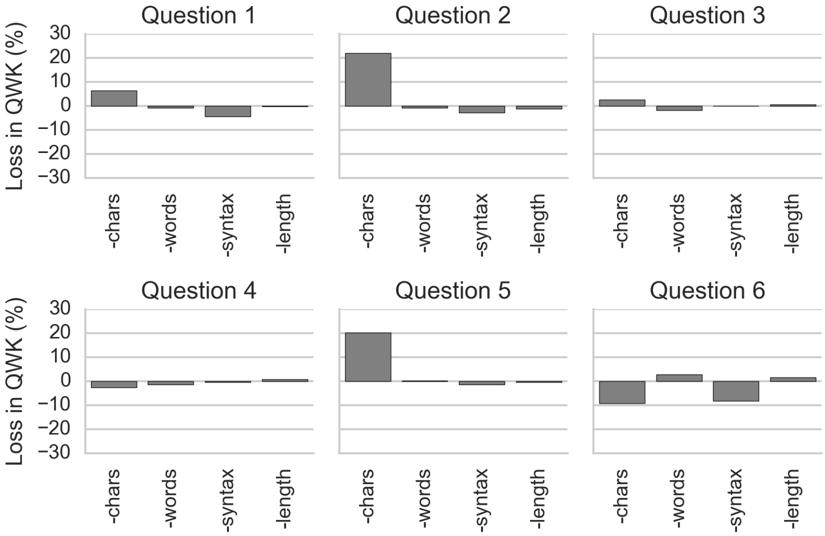
- lowercased word n-grams (n=1,2), including punctuation
- lowercased character n-grams (n=2,3,4,5)
- syntactic dependency triples computed using the ZPar parser (Zhang and Clark, 2011))
- length bins (specifically, whether the log of 1 plus the number of characters in the response, rounded down to the nearest integer, equals x, for all possible x from the training set)

A salient characteristic of this test and its constituent questions, as described by its designers, is that they measure content knowledge from prospective teachers, but not writing proficiency. There is a separate test that measures the writing proficiency of prospective teachers, that is required for all test-takers taking the music test.

In order to empirically confirm the minimal impact of writing proficiency, we build a second automated scoring model for writing proficiency using features inspired by Attali and Burstein (2006) and train it on the responses written by the same population of test-takers for the general writing proficiency test (*not* the music teaching proficiency test). Note that this model is generic, i.e., not question specific. We then use this trained model to assign scores to the responses from the music teaching proficiency test. A low agreement of these proficiency scores with the H1 scores assigned to the music questions should be sufficient evidence to indicate that the writing proficiency of the test-takers is not an important factor. There will obviously be some agreement because good writers

Q	N	QWK			Adjacent Agreement			Exact Agreement		
		H1-H2	H1-WP	H1-CS	H1-H2	H1-WP	H1-CS	H1-H2	H1-WP	H1-CS
1	1162	.702	.306	.566	.937	.817	.958	.705	.389	.470
2	1160	.764	.217	.570	.988	.892	1.00	.846	.434	.714
3	1154	.695	.245	.515	.955	.809	.945	.573	.358	.476
4	1272	.647	.253	.500	.959	.854	.962	.665	.398	.530
5	1270	.757	.089	.619	.983	.840	.994	.810	.330	.680
6	1267	.669	.196	.426	.950	.831	.928	.597	.359	.464

(a) Scoring performance on the test set for the six questions. N indicates the total number of responses available for each question, with 70% used to train the content scoring model, the H1-H2 columns denote the agreements between the two experts, the H1-WP columns denote the agreements between the H1 scores and those assigned by the generic writing proficiency model, and the H1-CS columns denote the agreements between the H1 scores and the question-specific content scoring model.



(b) The impact of ablating each of the four feature types on the overall scoring performance on the test set (*chars* = character n-grams, *words* = word n-grams, *syntax* = dependency triples, and *length* = log length features). Values > 0 indicate loss in performance when the feature is ablated and vice versa.

Figure 3: Scoring performance and ablation results.

are also likely to be better students.

5 Results

Figure 3a shows the performance of our content scoring model on the test set for all six questions (the H1-CS columns). We present three different metrics that measure the agreement of our model's predictions with the H1 scores. Although quadratic weighted kappa (QWK) is generally the standard metric of performance for short-answer scoring, we also compute the exact as well as adjacent agreement of the predictions with the H1 scores. The exact agreement shows the rate at which our model and H1 awarded the same score to a response. The adjacent agreement shows the rate at which scores given by our model and H1 were no more than one score point apart (e.g., the model assigned a score of 2 and the human rater assigned a score of 1 or 3). All three metrics were computed after rounding the raw predictions obtained from the SVR.

As an upper bound on automatic scoring performance, we also present the same agreement metrics between the H1 scores and the scores assigned by the second human expert (H2).

The table also includes the same agreement metrics for the predictions made by the generic writing proficiency model (the H1-WP columns). As expected, its performance is significantly worse than our content scoring model. This empirically confirms that the writing proficiency of the test-taker is not a factor in the human expert's assessment of their music teaching proficiency.

6 Discussion

Our model's predictions have high adjacent agreement with H1 scores. In fact, many adjacent agreement values are higher than the corresponding H1-H2 values. However, the exact agreement and QWK values are quite a bit lower than their H1-H2 counterparts. These observations tell us that although our content scoring model often predicts scores within 1 score point of the H1 score, it also either over-predicts or under-predicts the H1 score by more than 1 score point more often than H2 does.

Further spot-checking of sample responses in the training data indicated that sometimes it was possible that there was more than one correct answer to a question. For example, in the sample question from Section 2, it could be possible that there is more than one challenging aspect of the piece. As long as the test-taker articulates a valid challenge, along with an appropriate rehearsal technique, it is possible to obtain a score of 3. In situations where there is limited training data available, and not all valid challenging aspects have been sufficiently represented, for example, this may cause problems for automated scoring models. We cannot say with any certainty whether that caused the human-machine QWK scores to be lower than the corresponding human-human scores in our experiments, but it is an avenue of research that we intend to explore in future work.

We also wanted to examine how much each of the individual feature types contributes to the model's performance. To do so, we ablated each of the four feature types one at a time and re-ran the scoring model on the test set. Figure 3b shows the percentage loss in overall QWK for each of the six questions as we ablate each feature type. A value above zero indicates that removing a feature family led to a loss in performance and a value below zero indicates that removing a feature family actually led to an increase in performance.

We observe that including the *syntax* feature type almost always hurts the overall performance. At first, we hypothesized that this could be due to poor parser performance on these texts since they contain a lot of specialized musical terms (e.g., *glissando*, *embouchure* etc.) To confirm this hypothesis, we selected a few responses at random from the training data and looked at their dependency parses. Although we noticed some inaccuracies (e.g., *dotted* being interpreted as a verb in the phrase *rhythm dotted quarter note*), we did not find any evidence of significantly poor parsing performance. This means that the parsing feature representation itself seems to be deficient. We plan to experiment with other types of syntactic features in the future.

7 Conclusion

In this paper, we examined the feasibility of automatically scoring a unique content-based assessment - a test to measure the proficiency of teaching musical concepts to K-12 students. We first presented the characteristics that make the responses for this assessment different from almost all other previous work and then presented our approach to building an automated content scoring model. Our model performs moderately well on all six essay-style questions from the test but is prone to over- or under-predicting the true score by more than 1 point. As part of future work, we hope to explore the following in order to improve the model's performance:

- Increase the size of the training data to account for the relatively open-ended nature of the questions.
- Improve the representation of the syntactic features.
- Experiment with a hybrid approach (Sakaguchi et al., 2015) that combines our response-based approach with another approach that uses overlap with reference answers to assign scores.
- Experiment with some music- and instruction-specific features, including discourse/argumentation features.

Acknowledgments

We would like to thank Beata Beigman Klebanov, Matt Mulholland, Sunny Singh, Isaac Bejar, Seth Weiner, Kevin Cureton, Diane Bailey, and the anonymous BEA reviewers.

References

Enrique Alfonseca and Diana Pérez. 2004. Automatic assessment of open ended questions with a bleu-inspired algorithm and shallow nlp. In *Advances in Natural Language Processing*, pages 25–35. Springer.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater V. 2. *The Journal of Technology, Learning and Assessment*, 4(3):1–30.

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.

Myroslava 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. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 263–274, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 752–762, Portland, Oregon, USA, June. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Diana Pérez-Marín and Ismael Pascual-Nieto. 2011. Willow: a system to automatically assess students free-text answers by using a combination of shallow nlp techniques. *International Journal of Continuing Engineering Education and Life Long Learning*, 21(2-3):155–169.

Lakshmi Ramachandran, Jian Cheng, and Peter Foltz. 2015. Identifying patterns for short answer scoring using graph-based lexico-semantic text matching. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 97–106, Denver, Colorado, June. Association for Computational Linguistics.

Keisuke Sakaguchi, Michael Heilman, and Nitin Madnani. 2015. Effective feature integration for automated short answer scoring. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1049–1054, Denver, Colorado, May–June. Association for Computational Linguistics.

Alex J. Smola and Bernhard Schölkopf. 2004. A Tutorial on Support Vector Regression. *Statistics and Computing*, 14(3):199–222.

Jana Z Sukkarieh and Svetlana Stoyanchev. 2009. Automating model building in c-rater. In *Proceedings of the 2009 Workshop on Applied Textual Inference*, pages 61–69. Association for Computational Linguistics.

Jana Z Sukkarieh, Ali Mohammad-Djafari, Jean-Francois Bercher, and Pierre Bessiére. 2011. Using a maxent classifier for the automatic content scoring of free-text responses. In *AIP Conference Proceedings-American Institute of Physics*, volume 1305, page 41.

Yue Zhang and Stephen Clark. 2011. Syntactic Processing Using the Generalized Perceptron and Beam Search. *Computational linguistics*, 37(1):105–151.

Mengxiao Zhu, Ou L. Liu, Liyang Mao, and Amy Pallant. 2016. Use of Automated Scoring and Feedback in Online Interactive Earth Science Tasks. In *Proceedings of the 2016 IEEE Integrated STEM Education Conference*, Princeton, NJ.