Automated Scoring of 2023 NAEP Math Constructed-Response Items

Hong Jiao

Chandramani Yadav

Neil Shah

University of Maryland, College Park

This report summarizes technical details in developing automated scoring models for the NAEP math constructed-response items shared with us. We developed item specific models for each item. This report explains the model building process including data processing and model development. Further, this report summarizes the model performance in terms of quadratic weighted kappa (QWK) and standardized mean difference (SMD) for subgroup comparisons. The information presented in this report intends to summarize the accuracy of predicted scores from an automated scoring model compared with human ratings as well as the potential bias of the developed models for subgroups related to demographic variables including gender, race, LEP, IEP, and accommodation.

Model Building Process

To develop an automated scoring model for each NAEP math item, the whole model building process consists of two parts: feature extraction and model building. Before the process started, we pre-processed the raw text responses so that the decoding of the text data would be meaningful. Our modeling purely utilized information from student raw text responses. The quantification of the text response data was fulfilled via Natural Language Processing built in Bidirectional Encoder Representations from Transformers (BERT) and extended BERT models. We explored different approaches to model building via deep learning algorithms, more specifically BERT, extended BERT models and the ensemble methods of multiple deep learning models including BERT and LSTM and CNN. Once we identified the best-performing model for each item, we reported SMD for subgroups for potential bias evaluation based on the best performing model for an item.

Text Pre-Processing

The text strings in item responses may contain noise symbols or formatting. We pre-processed the text response strings before analyses were conducted. More specifically, the following cleaning steps were implemented for all training, validation, and test data. First, we turned all upper case letters into lower case in any response string. Then, hyperlinks such as 'https?:\/\/.*[\r\n]*' or 'http?:\/\/.*[\r\n]*' and hashtag signs # were removed. Some symbols were converted into meaningful words or symbols. These included converting '&amp;?' into 'and', converting '&lt;' into '<', and converting '&gt;' into '>'. Further, nonsense symbols like "(?:\@)\w+", "@", "\(", "\)", r'[:"#$%&*+,-/:;<=>@\\^_`{|}~]+' were removed. In addition, some non-ascii characters were identified and removed as well. Some punctuations were converted into clean formats including converting '[!]+' into '!', converting '[?]+' into '?', and converting '[.]+' into '.'. Further, we expanded the contractions like ‘dont’ to ‘do not’, ‘I'll’ to ‘I will’ using the contractions library.

Data Processing and Feature Extraction

A series of BERT models were explored in developing automated scoring models for each NAEP math constructed-response item with a score of 2 or 3. These models include BERT (Devlin et al., 2019), distilBERT (Sanh et al., 2020), DeBERTa-base (He et al., 2021), ELECTRA-base (Clark et al., 2020), and MathBERT (Shen et al., 2021). BERT is used to illustrate the steps for data processing and model development as follows. In general, data processing includes tokenization, padding, masking, and ultimately feature extractions.

1. Tokenization: using each BERT model’s tokenizer, sentences were first tokenized, broken up into words and subwords in the BERT format.
2. Padding: tokenized sentences of different lengths were all padded to the same size. Each response string was represented in the padded variables.
3. Masking: using attention_mask, to ignore the added padding in processing its input.
4. Feature extraction with deep learning: via each BERT model with 32 batch sizes and 3 epochs, the last_hidden_states were the results of the text processing and used as features for model building.

The processing of the text responses from each student and feature extraction were implemented within each model with the default settings. The same methods for processing of text data and feature extraction were conducted on training data, validation data, and the test data. The total number of tokens used as features for model building varied from one item to another depending on the maximum length of responses in an item. It varied from 100 to 512.

Automated Scoring Classifier Development

Once the features were extracted for the training and validation data, we developed an automated scoring classifier using multiple deep learning models. The explored models included BERT, MathBERT, distilBERT, DeBERTa, and ELECTRA. Among the 10 items, class imbalance could be an issue for at least two items: item3-VH266015 and item6-VH271613 as demonstrated in the sample sizes in Table 1 for each score category. Thus, we also applied the nlpaug package to rebalance the classes. In total, 7 models were experimented with in identifying the best performing model, namely BERT-Base, MathBERT with ANN plus nlpaug, MathBERT plus LSTM+CNN with nlpau, distilBERT, DeBERTa-Base, ELECTRA-Base, and ELECTRA-Base plus nlpaug. Though we experimented with more models including ALBERT (Lan et al., 2020) and RoBERTa (Liu et al. 2019), we did not report the results for these two models due to their relatively lower performance compared with other models.

Table 1
Sample Sizes for Training/Validation and Test Data and Sample Sizes for Each Score Category in the Training/Validation Dataset

Item #	Item ID	Sample size-Test Data	Sample Size-Training/Validation	Sample Size-Score 1	Sample Size-Score 2	Sample Size-Score 3
Item 1	VH134067	4,483	40,343	27,933	12,410	*
Item 2	VH139380	2,018	18,157	3,117	7,442	7,598
Item 3	VH266015	1,692	15,228	12,466	2,743	19
Item 4	VH266510	4,296	38,667	29,939	1,113	7,615
Item 5	VH269384	1,758	15,819	12,638	994	2,187
Item 6	VH271613	3,094	27,848	26,710	279	859
Item 7	VH302907	4,241	38,173	30,843	7,330	*
Item 8	VH304954	2,743	24,686	8,975	10,741	4,970
Item 9	VH507804	1,797	16,174	12,488	1,218	2,468
Item 10	VH525628	1,808	16,275	11,659	3,094	1,522

We used 80% vs 20% split for model training and validation. Once a model was trained, model performance was evaluated using the validation data. Our model performance was evaluated in two steps. First, we used the quadratic weighted kappa (QWK) to identify the best-performing model for each item. The bias was evaluated in terms of the standardized mean difference between the predicted scores from our best-performing automated scorer and the scores assigned by human raters.

Model Performance Evaluation

Table 2 summarizes QWKs for scoring the validation samples. In general, the QWK for our automated scorers ranged from 0.6 to 0.943. Given a QWK value larger than 0.7 is considered as a good rating and a value ranging from 0.4 to 0.7 as fair, our best-performing automated scorer for each item performed well with the highest QWK of 0.943 for two items and the lowest QWK of 0.684 for one item. Among the 10 items, the QWK for 3 items was around 0.94, another 3 items around 0.85, 2 items around 0.8, 1 item over 0.73, and 1 item over 0.68.

As presented in Table 2, no model performed the best consistently across all items in terms of QWK. ELECTRA-Base model performed the best on five items: item1-VH134067, item4-VH266510, item5-VH269384, item7-VH302907, and item9-VH507804. ELECTRA-Base model plus nlpaug for class imbalance performed the best on item3-VH226015. distilBERT performed the best on three items: item 2-VH139380, item8-CH304954, and item10-VH525628 though the three items had the lower QWK. MathBERT+LSTM+CNN+nlpaug performed the best on item6-VH271613. Item 3 and item 6 both had the class imbalance issue, the use of nlpaug for class rebalancing improved the model performance. In general, the ELECTRA models performed best on six items while the distilBERT model performed the best on three items and MathBERT performed well on one item. The better performance of the ELECTRA model is consistent with the evaluation of model performance by Phung (2021) who rank ordered the model performance as BERT, RoBERTa, ALBERT, ELECTRA, DeBERTa and ConvBERT with BERT as the reference base model to the best by ConvBERT and DeBERTa.

Table 2
Quadratic Weighted Kappa for the Developed Automated Scoring Models for Each Item

Item	Item ID	distilBERT	DeBERTa-Base	ELECTRA-Base	ELECTRA -Base+nlpaug	BERT-Base+nlp aug	MathBERT +ANN +nlpaug	MathBERT +LSTM+CNN +nlpaug
Item 1	VH134067	0.933	0.720	**0.938**	0.906	0.704	0.784	0.770
Item 2	VH139380	0.802	0.620	0.788	0.777	0.643	0.708	0.703
Item 3	VH266015	0.832	0.770	0.857	**0.865**	0.777	0.854	0.860
Item 4	VH266510	0.816	0.780	**0.849**	0.833	0.815	0.844	0.847
Item 5	VH269384	0.827	0.790	**0.860**	0.827	0.720	0.785	0.802
Item 6	VH271613	0.793	0.770	0.758	0.773	0.757	0.782	**0.796**
Item 7	VH302907	0.933	0.770	**0.943**	0.940	0.779	0.814	0.817
Item 8	VH304954	0.721	0.430	0.705	0.670	0.408	0.481	0.490
Item 9	VH507804	0.904	0.870	**0.943**	0.933	0.862	0.890	0.885
Item 10	VH525628	**0.684**	0.610	0.665	0.644	0.624	0.609	0.626

For the items with QWK smaller than or close to 0.80, namely items 2, 6, 8, and 10, we experimented with ConvBERT (Jiang et al., 2021). There was no improvement in QWK for items 6 and 10, but QWK for item 8 using ConvBERT increased to 0.734 from 0.721 using the distilBERT model. QWK for item 2 increased to 0.820 from 0.802 with 80-20 split but SMDs were slightly higher than 0.10 for a large number of subgroups. With 90-10 split for item 2, SMDs reduced for majority of the subgroups and QWK still increased to 0.817. So the results reported for item 2 were from 90-10-split. Table 3 summarizes the highest QWK values for each item based on the best-performing model.

Table 3
The Best-Performing Model for Each Item and Its Quadratic Weighted Kappa Compared with the Target

Item	Item ID	Human-Human QWK	Targeted QWK	Difference	Best QWK	Best Model
Item 1	VH134067	0.966	0.916	0.022	0.938	ELECTRA-Base
Item 2	VH139380	0.981	0.931	-0.114	0.817	ConvBERT
Item 3	VH266015	0.910	0.860	0.005	0.865	ELECTRA-Base +nlpaug
Item 4	VH266510	0.933	0.883	-0.034	0.849	ELECTRA-Base
Item 5	VH269384	0.948	0.898	-0.038	0.860	ELECTRA-Base
Item 6	VH271613	0.946	0.896	-0.100	0.796	MathBERT +LSTM+CNN+nlpaug
Item 7	VH302907	0.980	0.930	0.013	0.943	ELECTRA-Base
Item 8	VH304954	0.984	0.934	-0.200	0.734	ConvBERT
Item 9	VH507804	0.992	0.942	0.001	0.943	ELECTRA-Base
Item 10	VH525628	0.956	0.906	-0.222	0.684	distilBERT

Note: Difference= Difference between the expected and predicted QWK

To examine potential bias in the trained automated scorer for each item, standardized mean differences between the predicted scores based on the best-performing model trained for each item and the human scores on the validation sample were computed. Such differences were computed for five demographic variables, namely, gender, accommodation status, LEP, IEP, and race as summarized in Table 4. In general, SMDs for Pacific Islanders and American Indian were larger than 0.10 on seven and four items respectively. The red-colored cells indicate that SMDs were larger than 0.10. Hypothesis testing (paired t-test) was conducted. P-values are presented in parenthesis for each subgroup comparison. * indicates significant differences at the alpha level of 0.05. The largest SMD of 1.0037 was for item 6 for the LEP group. Most SMDs larger than 0.10 were not significant differences. One, two and two significant differences were observed for items 6, 8, and 10 where these three items had QWK lower than 0.80.

Table 4
SMD for Subgroups of Gender, Race, LEP, IEP, and Accommodation and Significant Difference

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Item ID	VH134067	VH139380	VH266015	VH266510	VH269384	VH271613	VH302907	VH304954	VH507804	VH525628
Overall	0.0289	0.0521	-0.0284	0.0358	0.0025	0.0662	0.0334	0.0740	0.0265	0.0085
Male	0.0154	0.0580	-0.0401	0.0327	-0.0246	0.0653	0.0326	0.0485	0.0219	-0.0239
Female	0.0441	0.0463	-0.0172	0.0392	0.0278	0.0671	0.0343	0.1024* (<.001)	0.0315	0.0417
Accommodated	0.0327	0.0693	-0.0755	0.0888	0.0626	0.0000	0.0839	*	-0.0276	-0.0568
Non-Accommodated	0.0291	0.0511	-0.0257	0.0348	0.0000	0.0672	0.0310	*	0.0297	0.0142
LEP	0.0000	0.1130 (0.103)	0.0526	0.0489	0.0471	1.0037* (0.029)	0.1396 (0.083)	0.0521	-0.0183	-0.3921* (0.032)
Non LEP	0.0313	0.0456	-0.0320	0.0359	0.0007	0.0616	0.0318	0.0770	0.0286	0.0184
IEP	0.0000	0.0557	-0.1865 (0.158)	0.0341	0.0000	-0.0878	0.0936	0.0694	0.0000	-0.0282
Non IEP	0.0318	0.0545	-0.0208	0.0364	0.0027	0.0712	0.0302	0.0762	0.0292	0.0118
White	0.0228	0.0376	-0.0111	0.0386	-0.0138	0.0720	0.0265	0.0669	0.0311	0.0684
Black	0.0505	0.0887	0.0000	0.0295	0.0713	0.1225	0.0507	0.2246* (<.001)	0.0600	-0.0899
Hispanic	0.0282	0.0486	-0.0889	0.0054	0.0402	0.1214	0.0554	0.0338	0.0327	-0.1071 (0.084)
Asian	0.0547	0.0511	-0.1690 (0.103)	0.0806	0.0093	0.0251	0.0402	-0.0523	-0.0111	-0.0236
American Indian	-0.0399	0.0000	-0.3626 (0.159)	0.0847	-0.0627	0.0000	0.1016 (0.319)	0.0288	-0.1698 (0.260)	-0.3901* (0.045)
Pacific Islander	0.0966	-0.5217 (0.172)	0.6849 (0.327)	0.1340 (0.103)	0.0000	-0.1569 (0.322)	0.3202 (0.159)	0.0000	-0.1942 (0.329)	0.1250 (0.714)
Other Race	0.0397	0.1655 (0.88)	0.1542 (0.158)	0.0563	-0.0603	0.0758	-0.0332	0.0985	0.0000	-0.0599

Note: the numbers inside the parenthesis are p-values for hypothesis testing. * indicates significant difference at the alpha level of 0.05.

To better understand SMDs, the sample sizes for each subgroup are reported in Table 5. In general, the sample size for Pacific Islanders was always the smallest while that for American Indian was the second smallest. This might be related to the relatively larger SMDs for these two groups though almost all differences were not significant except for one item.

Table 5
Sample Sizes for Subgroups of Gender, Race, LEP, IEP, and Accommodation

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Item ID	VH134067	VH139380	VH266015	VH266510	VH269384	VH271613	VH302907	VH304954	VH507804	VH525628
Overall	8069	1816	3046	7734	3164	5570	7635	4938	3235	3255
Male	4200	925	1520	3860	1570	2898	3905	2550	1634	1705
Female	3866	891	1525	3874	1594	2671	3730	2384	1600	1550
Accommodated	920	245	303	685	394	721	1284	*	502	580
Non-Accommodated	7147	1571	2743	7049	2769	4848	6351	*	2733	2675
LEP	808	184	178	427	328	611	480	449	368	216
Non LEP	7256	1630	2866	7307	2833	4958	7153	4487	2865	3039
IEP	956	255	300	741	386	661	1335	557	497	565
Non IEP	7109	1560	2745	6993	2777	4908	6298	4377	2735	2690
White	3991	895	1578	4108	1551	2837	4064	2544	1587	1692
Black	1394	305	481	1194	538	866	1261	811	563	547
Hispanic	1745	381	652	1610	686	1206	1539	1090	732	643
Asian	365	83	139	364	132	262	330	178	111	148
American Indian	166	54	65	152	75	103	154	85	77	88
Pacific Islander	68	7	26	58	32	49	58	38	21	23
Other Race	338	91	105	248	150	246	229	189	144	114

The patterns of model performance in terms of QWK and SMD were not consistent for some items. The three top-performing items 1, 7, and 9 with QWK around 0.94 still yielded a couple of SMD larger than 0.10 except item 1 with all SMD smaller than 0.10. Items 2, 3, 4, and 5 had a QWK larger than 0.8, but smaller than 0.9. Items 3, 4, and 5 all had QWK around 0.85. No SMD was larger than 0.10 for item 5, one SMD larger than 0.10 for item 4, and five SMDs were larger than 0.10 for item 3. On the other hand, item 2 with a QWK around 0.82 yielded SMD larger than 0.10 for only three subgroups. Items 6 and 8 had QWK around 0.75. Item 6 only had two SMDs larger than 0.10 for LEP (significant) and Pacific Islander groups (not significant). On the other hand, item 8 yielded SMDs larger than 0.10 for two groups: female and Black, both were significant. Though item 10 yielded the lowest QWK of around 0.68, only four SMDs were larger than 0.10 with two significant.

To better understand the model performance, we related the QWK values with the metadata of each item as presented in Table 6. There were no systematic patterns regarding sample size, cognitive complexity, grade level, and maximum scores. However, among the three items whose QWK was below 0.80, two of them were items on number properties and operations. Further exploration can be conducted to better understand the features of the item response strings.

Table 6
Metadata for Each Item and its QWK

Item #	Prompt ID	Max Score	Grade	Cognitive Complexity	Sample Size	Content Domain	Best-Performer QWK
Item 1	VH134067	2	4	Moderate	40,343	Algebra	0.938
Item 2	VH139380	3	4	Low	18,157	Algebra	0.817
Item 3	VH266015	3	8	Hard	15,228	NPO	0.865
Item 4	VH266510	3	8	Moderate	38,667	Algebra	0.849
Item 5	VH269384	3	4	Moderate	15,819	Statistics	0.860
Item 6	VH271613	3	4	High	27,848	Algebra	0.796
Item 7	VH302907	2	8	High	38,173	Geometry	0.943
Item 8	VH304954	3	4	Moderate	24,686	NPO	0.734
Item 9	VH507804	3	4	High	16,174	NPO	0.943
Item 10	VH525628	3	8	High	16,275	NPO	0.684

Note: NPO=number properties and operations; Statistics=data analysis, statistics, and probability

Scores were predicted for the test data using the best-performing model for each item, which is included in this submission.

Summary

In summary, we developed item-specific models for each of the grade 4 and 8 NAEP math items. In general, our best-performing model for each item yielded adequate QWK with values around or larger than 0.80. Though for the majority of items, SMD did not lead to concerns of potential bias among subgroups, SMDs for two items: item 8 and item 10, both on number properties and operations turned out to be large and the differences were significant for two subgroups respectively though the affected subgroups were not the same for the two items.

This report presents the technical details we followed in developing automated scoring models including pre-processing of the data, data processing feature extraction, and model development. In general, the interpretability is always a challenge in the application of deep learning algorithms. Our exploration focused on the BERT-based model and extensions. This may increase difficulty in the interpretability of the results and understanding the working of each best-performing automated scorer.

References

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). *ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators* (arXiv:2003.10555). arXiv. https://doi.org/10.48550/arXiv.2003.10555

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

He, P., Liu, X., Gao, J., & Chen, W. (2021). *DeBERTa: Decoding-enhanced BERT with Disentangled Attention* (arXiv:2006.03654). arXiv. https://doi.org/10.48550/arXiv.2006.03654

Jiang, Z., Yu, W., Zhou, D., Chen, Y.,

Jiang, Z., Yu, W., Zhou, D., Chen, Y., Feng, Y., & Yan, S. (2021). *ConvBERT: Improving BERT with Span-based Dynamic Convolution*. arxiv. https://arxiv.org/pdf/2008.02496.pdf

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). *ALBERT: A lite BERT for self-supervised learning of language representations*. Paper presented at ICLR. https://openreview.net/forum?id=H1eA7AEtvS

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. https://arxiv.org/pdf/1907.11692.pdf

Phung, T., M. (2021). *A review of pre-trained language models: from BERT, RoBERTa, to ELECTRA, DeBERTa, BigBird, and more*. Retrieved form https://tungmphung.com/a-review-of-pre-trained-language-models-from-bert-roberta-to-electra-deberta-bigbird-and-more/

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter* (arXiv:1910.01108). arXiv. https://doi.org/10.48550/arXiv.1910.01108

Shen, J., Yamashita, M., Prihar, E., Heffernan, N., Wu, X., & Lee, D. (2021). *MathBERT: A Pre-trained Language Model for General NLP Tasks in Mathematics Education*. Preprint at https://arxiv.org/pdf/2106.07340.pdf