Text-Based Approaches to Item Alignment to Content Standards in Reading & Writing Tests

Yanbin Fu, Hong Jiao, Tianyi Zhou, Nan Zhang, Ming Li, Qingshu Xu, Sydney Peters, Robert W. Lissitz
University of Maryland, College Park

Abstract

Aligning test items to content standards is a critical step in test development to collect validity evidence based on content. Item alignment has typically been conducted by human experts, but this judgmental process can be subjective and time-consuming. This study investigated the performance of fine-tuned small language models (SLMs) for automated item alignment using data from a large-scale standardized reading and writing test for college admissions. Different SLMs were trained for both domain and skill alignment. The model performance was evaluated using precision, recall, accuracy, weighted F1 score, and Cohen's kappa on two test sets. The impact of input data types and training sample sizes was also explored. Results showed that including more textual inputs led to better performance gains than increasing sample size. For comparison, classic suprvised machine learning classifiers were trained on multilingual-E5 embeddings. Fine-tuned SLMs consistently outperformed these models, particularly for fine-grained skill alignment. To better understand model classifications, semantic similarity analyses including cosine similarity, Kullback-Leibler divergence of embedding distributions, and two-dimension projections of item embeddings revealed that certain skills in the two test datasets were semantically too close, providing evidence for the observed misclassification patterns.

1. Introduction

Item alignment is part of alignment defined as the consistency among assessments, content standards, and instructional practices (Smith & O'Day, 1990; Webb, 1997). The degree of item alignment to content standards is critical evidence for validity based on content. Item alignment is typically conducted manually by content experts. The process involves reviewing test items one by one and determining which content standards each item aims to measure. Experts rely on their subject-matter expertise and professional judgement to assess alignment. Thus, this approach has clear limitations. First, manual alignment is time-consuming and labor-intensive especially for large-scale assessments (Bier et al., 2019; Ding et al., 2025; Zhou & Ostrow, 2022). Second, reliance on expert judgement introduces subjectivity (Camilli, 2024; Khan et al., 2021). Third, as test items are designed to measure more complex domains and skills, incorporating multiple skills, domains or hierarchical label structures makes manual methods increasingly insufficient (Li et al., 2024).

To address these limitations, researchers started exploring using machine learning and natural language processing (NLP) techniques. These approaches aim to enhance consistency, reduce labor, and enable scalability in large-scale assessment (Qu et al., 2011). Broadly, automated item alignment methods can be classified into two categories: feature-based models and language model-based approaches. Feature-based methods can be further divided into two categories: linguistic feature-based models and embedding-based models.

Recently, advances in transformer-based language models have introduced new modeling approaches to automated item alignment. These include small language models (SLMs), such as BERT, RoBERTa, and DeBERTa, which are often

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 3: Coordinated Session Papers, pages 19–36

fine-tuned on labeled items to directly map item text to the content standards (e.g., Ding et al., 2025; Shen et al., 2021; Tan & Kim, 2024). Another emerging trend involves large language models (LLMs), such as GPT-4, which use prompting or fine-tuning strategies to classify or generate labels without additional training (Li et al., 2024; Liu et al., 2025; Moore et al., 2024).

2. Related Work

Automated item alignment is typically formulated as a classification task, where the goal is to assign items to predefined content standards based on item text. Early studies relied on feature-based models. In supervised or unsupervised classification tasks, test items were mapped to one or more content labels using classifiers such as support vector machines (SVM; Karlovcec et al., 2012; Yilmazel et al., 2007), Latent Dirichlet Allocation (LDA; Anderson et al., 2020), and XGBoost (Tian et al., 2022). For instance, Karlovcec et al. (2012) applied SVM and K-nearest neighbor (KNN) to classify ASSISTments math items into 106 content labels, while Pardos and Dabu (2017) used skip-gram and bag-of-words features for item alignment to 198 content labels. Extracted linguistic features included bag-of-words, TF-IDF, and keyword overlaps, which did not well capture contextual or sequential information.

With the rise of neural network models, convolutional neural networks (CNNs; Kim, 2014) and recurrent neural networks (RNNs; Schuster & Paliwal, 1997) were adopted. BiLSTM, a type of RNN, was particularly effective for sequence modeling. Sun et al. (2018) showed that BiLSTM outperformed classic methods (e.g., SVM) in English question alignment with an F1 score of 0.562 vs. 0.447. More approaches employed embeddings extracted from Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), or contextual embeddings from models like BERT (Devlin et al., 2019). For example, Tian et al. (2022) used Word2Vec embeddings and keyphrase features with XGBoost to align high school math items, outperforming baseline models such as VSM, SVM, NN, and LSTM.

SLMs such as BERT and RoBERTa have been applied in item alignment using fine-tuned methods. Shen et al. (2021) found that fine-tuned BERT outperformed both classic classifiers and BERT model without fine-tuning. Khan et al. (2021) developed the Catalog system to align items with the NGSS standards using BERT and GPT-based semantic similarity measures. Tan and Kim (2024) compared FastText+XGBoost, fine-tuned BERT-base/large, RoBERTa-large, and GPT-3.5 with prompting, reporting that RoBERTa-large consistently performed best. Similarly, Ding et al. (2025) proposed a RoBERTa-based model, which outperformed BiLSTM, BiGRU, and BERT in math item alignment.

LLMs like GPT-3.5 and GPT-4 have also been explored for item alignment via prompting. Wang et al. (2023) used GPT-4 to classify medical test items using zero- and few-shot prompts. Li et al. (2024) explored alignment as binary classification task, prompting LLMs with item text and candidate knowledge descriptions along with a self-reflection step that allow the model to re-evaluate and revise its initial prediction. Their results showed that GPT-4 performed best, achieving over 90% accuracy. Moore et al. (2024) used GPT-4 to directly generate knowledge components, simulating expert annotation and even constructing hierarchical ontologies.

In summary, feature-based models extract linguistic features or use embeddings as features but often lack task adaptation. Fine-tuned SLMs, though less explored, offer an efficient middle ground between classic machine learning models and costly LLMs, with less privacy concern and better scalability for large-scale assessment contexts.

To address gaps in the literature on automated item alignment in large-scale educational assessment, this study investigates how SLMs can be fine-tuned for item content

alignment in large-scale reading and writing assessments. Specifically, this study addresses the following research questions:

1. How do sample size and input data type affect the item alignment accuracy?
2. How do different SLMs perform in aligning test items to skill and domain categories?
3. Where do misclassifications occur?

3. Methods

3.1 Data

This study used 1270 items from the SAT Reading and Writing (RW) section, with 80% for training and 20% for testing. Additionally, 1052 items from the PSAT 8/9 RW section were used as an external test set to evaluate fine-tuned models' generalizability. Each item includeed a prompt, a question, four answer options, the correct answer or key, and a rationale explaining both correct and incorrect answers. Some items also contain graphs or tables, which were converted into text descriptions and LaTeX respectively. Each item measures one of the 10 skills nested under 4 content domains including *Standard English Conventions*, *Information and Ideas*, *Expression of Ideas*, and *Craft and Structure*. Skill labels include *Boundaries*, *Form*, *Structure and Sense*, *Command of Evidence*, *Inferences*, *Central Ideas and Details*, *Transitions*, *Rhetorical Synthesis*, *Words in Context*, *Text Structure and Purpose*, and *Cross-Text Connections*.

3.2 Sample Size and Input Data

To investigate the impact of sample size and input data on item alignment accuracy, the study experimented with different sample sizes and input data in the training dataset. BERT-base was first used for such exploration. Specifically, this study first sampled 500, 750, and 1000 items from the full 1270 dataset. Each subset was further split into training and test datasets using a ratio of 80% vs 20%. Their training datasets contained 400, 600, and 800 items respectively. The models' performance was evaluated on test sets. Nine input data types were experimented as listed below:

1. Prompt only
2. Prompt+table+figure
3. Prompt+table+figure+options
4. Prompt+table+figure+options+key
5. Prompt+table+figure+options+key+rationale
6. Prompt+table+figure+question
7. Prompt+table+figure+question+options
8. Prompt+table+figure+question+options+key
9. Prompt+table+figure+question+options+key +rationale

3.3 Models

To answer the second question about SLMs performance in item alignment, several SLMs were fine-tuned. This study explored both SLM-based modeling approaches and embedding-based classic supervised machine learning models. The 12 fine-tuned SLMs include BERT-base, BERT-large (Devlin et al., 2019), ALBERT-base (Lan et al., 2019), DistilBERT-base (Sanh et al., 2019), All-DistilRoBERTa (Liu et al., 2019; Sanh et al., 2019), ELECTRA-small, ELECTRA-base (Clark et al., 2020), RoBERTa-base, RoBERTa-large (Liu et al., 2019), DeBERTa-base (He et al., 2020), DeBERTa-large (He et al., 2021), and ConvBERT (Jiang et al., 2020).

For comparison, embeddings from multilingual-E5-large-instruct model were extracted using the CLS token and used to train supervised machine learning models including logistic regression, SVM, Naive Bayes, Random Forest, Gradient Boosting, XGBoost, LightGBM, MLP, and KNN.

3.4 Model Fine-Tuning

Prior to setting up the training configuration, this study conducted a series of exploratory experiments to evaluate the effects of different hyperparameter settings. Specifically, this study compared multiple learning rates (1e-5, 2e-5, and 3e-5), warm-up ratio (0 and 0.1), learning rate scheduler (linear and cosine), and checkpoints (epoch-wise and step-wise). Based on model performance with different settings, the following configuration was selected for all models. That is, models were trained with 15 epochs using the AdamW optimizer, a learning rate of 2e-5, a batch

size of 8, and a linear learning rate scheduler with a warmup ratio of 0.1. Each SLM was fine-tuned separately for the domain and skill alignment. Item input texts were tokenized using the tokenizer of each SLM and truncated to a maximum length of 512 tokens. The model performance was evaluated in terms of accuracy, recall, precision, weighted F1 score, and Cohen's kappa coefficient on both the SAT test dataset and the PSAT items.

3.5 Exploration for Misclassification

To understand the underlying causes of model misclassification, this study used a range of embedding-based analytical techniques. First, this study calculated all-pairwise cosine similarity between the selected skill groups with high rates of observed misclassification to quantify their semantic proximity in the embedding space. Second, To visualize the structure of the embeddings, this study applied three common dimensionality reduction techniques, including principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and isometric mapping (ISOMAP), to project the item embeddings from the best performing models into a two dimensional space for the clustering patterns. Third, KL divergence was calculated between skill-specific embedding distributions. Lower KL scores suggest semantically similarity.

4. Results

4.1 Impact of Sample Size and Input Data

This study examined how input data and sample size affected item alignment accuracy using the BERT-base model. As shown in Table A.1 and A.2 in appendix, input data had a more substantial impact than sample size on both skill and domain alignment performance. Across all sample sizes, models trained with minimal inputs of "prompt_only" consistently yielded the lowest performance, while including more item components such as options, keys, rationales, and question improved model performance. For instance, in the skill alignment task with 400 training samples, weighted F1 score increased from 0.664 with "prompt_only" to 0.919 with all input data. However, the accuracy increase was not monotonic along with adding more input data. For example, when 400 items were used for training, adding the rationale led to decreased weighted F1 from 0.981 to 0.935.

It is worthy of note, adding question resulted in a sharp jump in alignment accuracy. For example, when 400 items were used for training, weighted F1 score for skill alignment increased from 0.664 with "prompt_only" to 0.893 with "prompt_table_figure_qtext". This dramatic increase was due to that many items in the same domain such as "Standard English Conventions" shared nearly identical question templates like "Which choice completes the text so that it conforms to the conventions of Standard English?" These question templates were likely to act as shortcut features, allowing models to memorize superficial patterns rather than learn the semantic relationship between content and skill or domain labels. To mitigate this issue, all questions was removed from the input data.

In contrast, increasing the training sample size from 400 to 800 yielded modest improvement, particularly when compared with the increase achieved through adding input data. For example, for skill alignment with "prompt_only," weighted F1 score improved from 0.664 for a sample size of 400 to 0.787 for a sample size of 600, whereas the same level of performance increase could be surpassed by adding more input data even with small sample sizes. A similar pattern was observed for domain alignment even though weighted F1 score was 0.919 with a sample size of 400 and "prompt_only" but F1 score increased to 0.927 with a sample size of 600 and all input data. These findings suggested that though larger training sample size increased accuracy, the more input data led to larger improvement in alignment accuracy more effectively.

4.2 The Impact of Hyper-Parameters for Fine-Tuning SLMs

To evaluate the effect of fine-tuning settings,

a full factorial experiment was conducted using BERT-base with different combinations of learning rate (1e-5, 2e-5, 5e-5), warm-up ratio (0.0, 0.1), learning rate scheduler (linear, cosine), and checkpoint strategy (epoch-wise, step-wise). The results showed that BERT-base model maintained strong performance across all hyper-parameter combinations. Weighted F1 scores, accuracy, and Cohen's kappa remained above 0.98 in nearly all cases, indicating a high degree of robustness to hyper-parameter choices.

4.3 Model Performance Comparison

Tables A.3 and A.4 in Appendix compared model performance on the SAT test set for skill and domain alignment. Across all metrics, fine-tuned SLMs significantly outperformed classical embedding-based classifiers. For skill alignment, ConvBERT and RoBERTa-large achieved perfect scores on all metrics, and even the worst performing ALBERT-base still performed well with weighted F1 of 0.943. Feature-based classifiers yielded lower performance, with weighted F1 scores ranging from 0.513 to 0.829. Among them, MLP showed the best performance. Domain alignment appeared to be an easier task, with most SLMs achieving nearly perfect results. Several models, including RoBERTa-large, ConvBERT, and DeBERTa-base, achieved perfect scores on all metrics. Feature-based classifiers also performed reasonably well, with weighted F1 scores generally above 0.84, indicating domain alignment task was easier.

The generalizability of fine-tuned SLMs was further tested on the PSAT dataset (Tables A.5 and A.6). While model performance dropped slightly compared to SAT test data, most models still performed well. For skill alignment, ELECTRA-base and RoBERTa-large remained the best performance with weighted F1 scores larger than 0.99, and DeBERTa-base and ALBERT-base performed well too with F1 score larger than 0.95. For domain alignment, DeBERTa-base performed best with all metrics having a value of 0.997. RoBERTa-base, RoBERTa-large also performed well with all metrics of 0.994. These findings suggest that models trained on SAT items can be generalized to PSAT item alignment when the same content framework are followed.

4.4 Exploration of Misclassification

Though the overall accuracy of aligning PSAT items was high using the model trained on SAT items, some skill-specific item alignment displayed high misclassification rate. Table A.7 presents F1 scores for skills on PSAT items. Several models, including BERT-base, BERT-large, ConvBERT, All-DistilRoBERTa, ELECTRA-small, RoBERTa-base, DeBERTa-large, and DistilBERT-base exhibited evident decrease in F1 scores on Skill 4 for *Inferences* and Skill 5 for *Central Ideas and Details*. Items for assessing these two Skills were often misclassified into Skill 8 for *Words in Context*.

To investigate misclassification, this study computed pairwise cosine similarities between iembeddings of items assessing Skills 4, 5, and 8 in SAT and PSAT. Results revealed high semantic similarity between Skill 4 and 8 with mean cosine similarity of 0.827 for SAT and 0.828 for PSAT and between Skill 5 and 8 with mean cosine similarity of 0.825 for SAT and 0.823 for PSAT.

Further, this study visualized the item-level embeddings using dimensionality reduction techniques, including PCA, t-SNE, and ISOMAP. The two-dimension projected embeddings for Skills 4 and 8, as well as Skills 5 and 8, showed considerable overlap across six plots. The four skill clusters occupied overlapping regions in the latent space, with no clear visual boundaries between them, indicating that the items shared highly similar semantic characteristics.

In addition, KL divergence was used to assess how PSAT Skills 4 and 5 align with each SAT skill in the embedding space. The results showed that SAT Skill 8 consistently exhibited low KL divergence (17.986 and 25.491) with the two PSAT skills, indicating the high semantic similarity. These results provide empirical evidence showing the semantic similarity between PSAT Skills 4/5 items and Skill 8 respectively where misclassification occurred.

5. Discussion and Conclusion

This study fine-tuned SLMs for automated item alignment in large-scale reading and writing assessments. Using SAT and PSAT data, items were aligned to both domains and skills, with skills nested within domains. The results demonstrated that fine-tuned SLMs substantially outperformed embedding-based classic machine learning models. Fine-tuned SLMs achieved high performance across all metrics, particularly in domain alignment. Even the weakest model, ALBERT-base, yielded weighted F1 score of 0.943. In contrast, embedding-based models trained on SLM yielded F1 scores ranging from 0.513 to 0.829, highlighting the superiority of end-to-end fine-tuning of SLMs.

More input data consistently outperformed the models trained with fewer input data. Increasing the sample size alone yielded relatively moderate improvements in model performance, especially when the input data were limited. However, the benefit of more input data was not monotonically increasing. With a sample size of 500, adding the rationale to the input data alongside the prompt, tables, figures, question, options, and key led to decreased performance. As sample size increased, this negative effect disappeared, suggesting an interaction between input data and sample size.

ELECTRA-base, RoBERTa-large, and DeBERTa-base demonstrated good generalizability on PSAT item alignment. Nevertheless, items measuring *Inferences* as well as *Central Ideas and Details* were frequently misclassified as *Words in Context*. Cosine Similarity and KL divergence analysis confirmed high overlapping in the embedding space across these skills, while two dimension projections using PCA, t-SNE, and ISOMAP further illustrated indistinct category boundaries.

Despite the promising results of SLMs in item content alignment demonstrated, this study has some limitations. First, items were all single-coded items. In some item content alignment, items may be double, triple, even multiple coded. Future research can explore more complex multi-coded item content alignment. Second, LLMs such as GPT-4 have shown promise in recent studies, they were not included in this study due to cost, transparency, and test security concerns. Future work may examine prompt-based LLMs alongside fine-tuned SLMs to assess their relative strengths in large-scale educational assessment programs.

In summary, this study evaluated multiple SLMs for automated item alignment to content standards. The investigation of the impact of sample size and input data types provided empirical evidence about these design factors in training SLMs for automated item alignment. The analyses related to misclassification errors help future studies to conduct quality control of any low performing cases. Though the current study used SAT and PSAT Reading and Writing items, the methods used for developing models for automated item alignment can be readily applied to state assessment programs when item alignment to content standards is needed.

References

Anderson, D., Rowley, B., Stegenga, S., Irvin, P. S., & Rosenberg, J. M. (2020). Evaluating content-related validity evidence using a text-based machine learning procedure. Educational Measurement: Issues and Practice, 39(4), 53-64.

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. Educational Measurement: Issues and Practice, 22(3), 21–29.

Bier, N., Moore, S., & Van Velsen, M. (2019, March). Instrumenting courseware and leveraging data with the Open Learning Initiative (OLI). In Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK19)

Butterfuss, R., & Doran, H. (2025). An application of text embeddings to support alignment of educational content standards. Educational Measurement: Issues and Practice, 44(1), 73–83.

Camilli, G. (2024). An NLP crosswalk between the Common Core State Standards and NAEP item

specifications. arXiv preprint arXiv:2405.17284.

Christopherson, S. C., & Webb, N. L. (2020). Alignment Analysis of Two Forms of the SAT with the Arizona Academic Standards for English Language Arts Grades 11–12, Algebra 1, and Geometry. Wisconsin Center for Education Products and Services.

Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., & Hu, G. (2020). Revisiting pre-trained models for Chinese natural language processing. In T. Cohn, Y. He, & Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 657–668). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.58

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (pp. 4171–4186).

Ding, Z., Wang, X., Wu, Y., Cao, G., & Chen, L. (2025). Tagging knowledge concepts for math problems based on multi-label text classification. Expert Systems with Applications, 267, 126232.

Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., & Smith, N. (2020). Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. arXiv preprint arXiv:2002.06305.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Psychology Press.

He, P., Gao, J., & Chen, W. (2021). Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. arXiv preprint arXiv:2111.09543.

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654.

Herman, J. L., Webb, N. M., & Zuniga, S. (2003). Alignment and college admissions: The match of expectations, assessments, and educator perspectives. Center for the Study of Evaluation, CRESST, UCLA.

Huang, T., Hu, S., Yang, H., Geng, J., Liu, S., Zhang, H., & Yang, Z. (2023). PQSCT: Pseudo-Siamese BERT for concept tagging with both questions and solutions. IEEE Transactions on Learning Technologies, 16(5), 831–846. https://doi.org/10.1109/TLT.2023.3275707

Nemeth, Y., Michaels, H., Wiley, C., & Chen, J. (2016). Delaware system of student assessment and Maine comprehensive assessment system: SAT alignment to the Common Core State Standards. Human Resources Research Organization.

Jiang, Z. H., Yu, W., Zhou, D., Chen, Y., Feng, J., & Yan, S. (2020). Convbert: Improving bert with span-based dynamic convolution. Advances in Neural Information Processing Systems, 33, 12837-12848.

Kane, M. (2006). Content-Related Validity Evidence in Test Development. In S. M. Downing & T. M. Haladyna (Eds.), Handbook of test development (pp. 131–153). Lawrence Erlbaum Associates.

Karlovčec, M., Córdova-Sánchez, M., & Pardos, Z. A. (2012). Knowledge component suggestion for untagged content in an intelligent tutoring system. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.), Intelligent tutoring systems: 11th International Conference, ITS 2012, Chania, Crete, Greece, June 14–18, 2012. Proceedings (Lecture Notes in Computer Science, Vol. 7315, pp. 195–200). Springer. https://doi.org/10.1007/978-3-642-30950-2_25

Khan, S., Rosaler, J., Hamer, J., & Almeida, T. (2021). Catalog: An educational content tagging system. In Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021). International Educational Data Mining Society.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1746–1751). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1181

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.

Li, H., Xu, T., Tang, J., & Wen, Q. (2024). Automate knowledge concept tagging on math questions with LLMs. arXiv preprint arXiv:2403.17281.

Lima, P. S. N., Ambrosio, A. P., Felix, I., Brancher, J. D., & de Carvalho, D. T. (2018). Content analysis of student assessment exams. In 2018 IEEE Frontiers in Education Conference (FIE) (pp. 1–9). IEEE. https://doi.org/10.1109/FIE.2018.8659169

Liu, N., Sonkar, S., Basu Mallick, D., Baraniuk, R., & Chen, Z. (2025). Atomic learning objectives and LLMs labeling: A high-resolution approach for physics education. In Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25) (pp. 620–630). Association for Computing Machinery. https://doi.org/10.1145/3706468.3706550

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. Review of Educational Research, 79(4), 1332–1361.

McCormick, C., & Geisinger, K. F. (2017). Alignment Study Full Report. Buros Center for Testing.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Moore, S., Schmucker, R., Mitchell, T., & Stamper, J. (2024). Automated generation and tagging of knowledge components from multiple-choice questions. In Proceedings of the Eleventh ACM Conference on Learning @ Scale (L@S '24) (pp. 122–133). Association for Computing Machinery. https://doi.org/10.1145/3657604.3662030

Mosbach, M., Andriushchenko, M., & Klakow, D. (2020). On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. arXiv preprint arXiv:2006.04884.

Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2022). MTEB: Massive text embedding benchmark. arXiv preprint arXiv:2210.07316.

Nemeth, Y., Michaels, H., Wiley, C., & Chen, J. (2016). Delaware System of Student Assessment and Maine Comprehensive Assessment System: SAT alignment to the Common Core State Standards – Final Report. Human Resources Research Organization.

Ozyurt, Y., Feuerriegel, S., & Sachan, M. (2025). Automated knowledge concept annotation and question representation learning for knowledge tracing. arXiv Preprint, arXiv:2410.01727. https://doi.org/10.48550/arXiv.2410.01727

Pardos, Z. A., & Dadu, A. (2017). Imputing KCs with representations of problem content and context. In Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP '17) (pp. 148–155). Association for Computing Machinery. https://doi.org/10.1145/3079628.3079689

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532–1543).

Peters, S., Zhang, N., Jiao, H., Li, M., Zhou, T., Lissitz, R., Fu, Y., & Xu, Q. (2025). Text-based approaches to item difficulty modeling in high-stakes assessments: A systematic review (MARC Research Report). University of Maryland.

Qu, B., Cong, G., Li, C., Sun, A., & Chen, H. (2012). An evaluation of classification models for question topic categorization. Journal of the American Society for Information Science and Technology, 63(5), 889-903.

Ramesh, R., Sasikumar, M., & Iyer, S. (2016). A software tool to measure the alignment of assessment instrument with a set of learning objectives of a course. In 2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT) (pp. 64–68). IEEE. https://doi.org/10.1109/ICALT.2016.10

Reimers, N., & Gurevych, I. (2021). all-distilroberta-v1 [Computer software]. Hugging Face. https://huggingface.co/sentence-transformers/all-distilroberta-v1

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11), 2673–2681.

Sebastiani, F. (2002). Machine learning in automated text

categorization. ACM computing surveys (CSUR), 34(1), 1-47.

Shen, J. T., Yamashita, M., Prihar, E., Heffernan, N., Wu, X., McGrew, S., & Lee, D. (2021). Classifying math knowledge components via task-adaptive pre-trained BERT. In Artificial Intelligence in Education: 22nd International Conference, AIED 2021, Utrecht, The Netherlands, June 14–18, 2021, Proceedings, Part I 22 (pp. 408-419). Springer International Publishing.

Smith, M. S., & O'Day, J. (1990). Systemic school reform. Journal of Education Policy, 5(5), 233–267. https://doi.org/10.1080/02680939008549074

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of documentation, 28(1), 11-21.

Sun, B., Zhu, Y., Xiao, Y., Xiao, R., & Wei, Y. (2018). Automatic question tagging with deep neural networks. IEEE Transactions on Learning Technologies, 12(1), 29-43.

Tan, C. S., & Kim, J. J. (2024). Automated Math Word Problem Knowledge Component Labeling and Recommendation. In International Conference in Methodologies and intelligent Systems for Techhnology Enhanced Learning (pp. 338-348). Cham: Springer Nature Switzerland.

Tian, Z., Flanagan, B., Dai, Y., & Ogata, H. (2022). Automated matching of exercises with knowledge components. In 30th International Conference on Computers in Education Conference Proceedings (pp. 24-32).

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). Multilingual e5 text embeddings: A technical report. arXiv preprint arXiv:2402.05672.

Wang, T., Stelter, K., Floyd, J., O'Neill, T., Hendrix, N., Bazemore, A., Rode, K., & Newton, W. (2023). Blueprinting the future: Automatic item categorization using hierarchical zero-shot and few-shot classifiers. arXiv. https://arxiv.org/abs/2312.03561

Webb, N. L. (1997). Criteria for alignment of expectations and assessments in mathematics and science education. Research Monograph No. 6.

Yilmazel, O., Balasubramanian, N., Harwell, S. C., Bailey, J., Diekema, A. R., & Liddy, E. D. (2007). Text categorization for aligning educational standards. In 2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07) (p. 73). IEEE. https://doi.org/10.1109/HICSS.2007.517

Yu, R., Das, S., Gurajada, S., Varshney, K., Raghavan, H., & Lastra-Anadon, C. (2021). A research framework for understanding education-occupation alignment with NLP techniques. In A. Field, S. Prabhumoye, M. Sap, Z. Jin, J. Zhao, & C. Brockett (Eds.), Proceedings of the 1st Workshop on NLP for Positive Impact (pp. 100–106). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.nlp4posimpact-1.11

Zhang, N., Jiao, H., Yadav, C., & Lissitz, R. (2025). Aligning SAT math to state math content standards: A systematic review (Technical report). Maryland Assessment Research Center, University of Maryland.

Zhou, Z., & Ostrow, K. S. (2022). Transformer-based automated content-standards alignment: A pilot study. In G. Meiselwitz (Ed.), HCI International 2022 – Late Breaking Papers: Interaction in New Media, Learning and Games (Vol. 13517, pp. 525–542). Springer. https://doi.org/10.1007/978-3-031-22131-6_39

Appendix

Table A.1

Performance of BERT-base Models across Sample Sizes and Input Data for Skill Alignment

Sample Sizes	Input Conditions	Accuracy	Precision	Recall	Weighted F1	Cohen's Kappa
400	prompt_only	0.700	0.690	0.700	0.664	0.662
	prompt_table_figure	0.810	0.813	0.810	0.801	0.786
	prompt_table_figure_options	0.900	0.904	0.900	0.897	0.886
	prompt_table_figure_options_key	0.880	0.886	0.880	0.876	0.864
	prompt_table_figure_options_key_rationale	0.920	0.926	0.920	0.919	0.909
	prompt_table_figure_qtext	0.890	0.915	0.890	0.893	0.876
	prompt_table_figure_qtext_options	1.000	1.000	1.000	1.000	1.000
	prompt_table_figure_qtext_options_key	0.980	0.984	0.980	0.981	0.977
	prompt_table_figure_qtext_options_key_rationale	0.940	0.970	0.940	0.935	0.932
600	prompt_only	0.787	0.796	0.787	0.787	0.760
	prompt_table_figure	0.767	0.795	0.767	0.754	0.738
	prompt_table_figure_options	0.880	0.876	0.880	0.871	0.865
	prompt_table_figure_options_key	0.900	0.911	0.900	0.898	0.887
	prompt_table_figure_options_key_rationale	0.933	0.948	0.933	0.932	0.925
	prompt_table_figure_qtext	0.947	0.948	0.947	0.947	0.940
	prompt_table_figure_qtext_options	0.993	0.994	0.993	0.993	0.992
	prompt_table_figure_qtext_options_key	0.980	0.980	0.980	0.980	0.977
	prompt_table_figure_qtext_options_key_rationale	0.980	0.982	0.980	0.980	0.977
800	prompt_only	0.800	0.817	0.800	0.798	0.777
	prompt_table_figure	0.815	0.812	0.815	0.811	0.793
	prompt_table_figure_options	0.865	0.887	0.865	0.871	0.849
	prompt_table_figure_options_key	0.890	0.915	0.890	0.896	0.877
	prompt_table_figure_options_key_rationale	0.850	0.883	0.850	0.855	0.832
	prompt_table_figure_qtext	0.950	0.950	0.950	0.950	0.944
	prompt_table_figure_qtext_options	0.990	0.990	0.990	0.990	0.989
	prompt_table_figure_qtext_options_key	0.995	0.995	0.995	0.995	0.994
	prompt_table_figure_qtext_options_key_rationale	0.995	0.995	0.995	0.995	0.994

Table A.2

Performance of BERT-base Models across Sample Sizes and Input Data for Domain Alignment

Sample Sizes	Input Conditions	Accuracy	Precision	Recall	Weighted F1	Cohen's Kappa
400	prompt_only	0.920	0.929	0.920	0.919	0.891
	prompt_table_figure	0.930	0.931	0.930	0.930	0.905
	prompt_table_figure_options	0.960	0.963	0.960	0.960	0.945

	prompt_table_figure_options_key	0.970	0.973	0.970	0.970	0.959
	prompt_table_figure_options_key_rationale	0.990	0.990	0.990	0.990	0.986
	prompt_table_figure_qtext	1.000	1.000	1.000	1.000	1.000
	prompt_table_figure_qtext_options	0.970	0.973	0.970	0.970	0.959
	prompt_table_figure_qtext_options_key	1.000	1.000	1.000	1.000	1.000
	prompt_table_figure_qtext_options_key_rationale	0.980	0.981	0.980	0.980	0.973
600	prompt_only	0.900	0.900	0.900	0.900	0.866
	prompt_table_figure	0.900	0.902	0.900	0.899	0.866
	prompt_table_figure_options	0.953	0.958	0.953	0.954	0.937
	prompt_table_figure_options_key	0.953	0.960	0.953	0.954	0.937
	prompt_table_figure_options_key_rationale	0.927	0.934	0.927	0.927	0.902
	prompt_table_figure_qtext	1.000	1.000	1.000	1.000	1.000
	prompt_table_figure_qtext_options	1.000	1.000	1.000	1.000	1.000
	prompt_table_figure_qtext_options_key	1.000	1.000	1.000	1.000	1.000
	prompt_table_figure_qtext_options_key_rationale	0.987	0.987	0.987	0.987	0.982
800	prompt_only	0.885	0.888	0.885	0.885	0.846
	prompt_table_figure	0.900	0.901	0.900	0.900	0.866
	prompt_table_figure_options	0.965	0.966	0.965	0.965	0.953
	prompt_table_figure_options_key	0.960	0.962	0.960	0.960	0.947
	prompt_table_figure_options_key_rationale	0.940	0.947	0.940	0.941	0.920
	prompt_table_figure_qtext	1.000	1.000	1.000	1.000	1.000
	prompt_table_figure_qtext_options	1.000	1.000	1.000	1.000	1.000
	prompt_table_figure_qtext_options_key	1.000	1.000	1.000	1.000	1.000
	prompt_table_figure_qtext_options_key_rationale	0.990	0.990	0.990	0.990	0.987

Table A.3

Model Performance on SAT Skill Alignment

Model	Precision	Recall	Accuracy	Weighted F1	Cohen's Kappa
BERT-base	0.996	0.996	0.996	0.996	0.996
BERT-large	0.989	0.988	0.988	0.988	0.987
ALBERT-base	0.949	0.945	0.945	0.943	0.938
ConvBERT	1.000	1.000	1.000	1.000	1.000
All-DistilRoBERTa	0.985	0.984	0.984	0.984	0.982
ELECTRA-base	0.992	0.992	0.992	0.992	0.991
ELECTRA-small	0.974	0.969	0.969	0.966	0.965
RoBERTa-base	0.996	0.996	0.996	0.996	0.996
RoBERTa-large	1.000	1.000	1.000	1.000	1.000
DeBERTa-base	0.985	0.984	0.984	0.984	0.982
DeBERTa-large	0.996	0.996	0.996	0.996	0.996
DistilBERT-base	0.992	0.992	0.992	0.992	0.991
Logistic Regression	0.538	0.646	0.646	0.563	0.593
SVM	0.642	0.701	0.701	0.643	0.658

Naive Bayes	0.764	0.744	0.744	0.749	0.713
Random Forest	0.591	0.610	0.571	0.513	0.554
Gradient Boosting	0.575	0.583	0.594	0.573	0.526
XGBoost	0.618	0.610	0.610	0.597	0.560
LightGBM	0.652	0.665	0.665	0.643	0.621
MLP	0.816	0.823	0.835	0.829	0.800
KNN	0.524	0.535	0.535	0.513	0.476

Table A.4

Model Performance on SAT Domain Alignment

Model	Precision	Recall	Accuracy	Weighted F1	Cohen's Kappa
BERT-base	0.996	0.996	0.996	0.996	0.995
BERT-large	0.996	0.996	0.996	0.996	0.995
ALBERT-base	0.967	0.965	0.965	0.965	0.952
ConvBERT	1.000	1.000	1.000	1.000	1.000
All-DistilRoBERTa	0.996	0.996	0.965	0.965	0.995
ELECTRA-base	0.996	0.996	0.996	0.996	0.995
ELECTRA-small	0.980	0.980	0.980	0.980	0.973
RoBERTa-base	1.000	1.000	1.000	1.000	1.000
RoBERTa-large	1.000	1.000	1.000	1.000	1.000
DeBERTa-base	1.000	1.000	1.000	1.000	1.000
DeBERTa-large	0.996	0.996	0.996	0.996	0.995
DistilBERT-base	0.992	0.992	0.992	0.992	0.989
Logistic Regression	0.879	0.878	0.878	0.878	0.834
SVM	0.901	0.894	0.894	0.894	0.857
Naive Bayes	0.839	0.827	0.827	0.827	0.767
Random Forest	0.812	0.807	0.783	0.781	0.735
Gradient Boosting	0.852	0.850	0.846	0.846	0.796
XGBoost	0.829	0.823	0.823	0.824	0.760
LightGBM	0.848	0.846	0.846	0.847	0.792
MLP	0.923	0.921	0.921	0.921	0.893
KNN	0.727	0.724	0.724	0.719	0.627

Table A.5

Model Performance on PSAT Skill Alignment

Model	Precision	Recall	Accuracy	Weighted F1	Cohen's Kappa
BERT-base	0.935	0.894	0.894	0.878	0.879
BERT-large	0.906	0.827	0.827	0.797	0.802
ALBERT-base	0.969	0.961	0.961	0.961	0.956
ConvBERT	0.902	0.887	0.887	0.870	0.871
All-DistilRoBERTa	0.931	0.907	0.907	0.887	0.895
ELECTRA-base	0.993	0.993	0.993	0.993	0.993
ELECTRA-small	0.744	0.760	0.760	0.722	0.728

RoBERTa-base	0.959	0.942	0.942	0.929	0.935
RoBERTa-large	0.994	0.994	0.994	0.994	0.994
DeBERTa-base	0.978	0.976	0.976	0.976	0.973
DeBERTa-large	0.927	0.894	0.894	0.868	0.879
DistilBERT-base	0.940	0.920	0.920	0.910	0.910
Logistic Regression	0.708	0.723	0.723	0.653	0.682
SVM	0.861	0.804	0.804	0.763	0.776
Naive Bayes	0.862	0.853	0.853	0.855	0.834
Random Forest	0.938	0.933	0.920	0.919	0.924
Gradient Boosting	0.881	0.879	0.883	0.882	0.864
XGBoost	0.917	0.914	0.914	0.914	0.903
LightGBM	0.938	0.937	0.937	0.936	0.929
MLP	0.963	0.963	0.961	0.961	0.958
KNN	0.695	0.695	0.695	0.687	0.655

Table A.6

Model Performance on PSAT Domain Alignment

Model	Precision	Recall	Accuracy	Weighted F1	Cohen's Kappa
BERT-base	0.947	0.934	0.934	0.934	0.912
BERT-large	0.986	0.985	0.985	0.985	0.980
ALBERT-base	0.892	0.820	0.820	0.803	0.762
ConvBERT	0.971	0.967	0.967	0.967	0.956
All-DistilRoBERTa	0.986	0.986	0.986	0.986	0.981
ELECTRA-base	0.928	0.904	0.904	0.902	0.872
ELECTRA-small	0.949	0.937	0.937	0.937	0.916
RoBERTa-base	0.994	0.994	0.994	0.994	0.992
RoBERTa-large	0.994	0.994	0.994	0.994	0.992
DeBERTa-base	0.997	0.997	0.997	0.997	0.996
DeBERTa-large	0.988	0.988	0.988	0.988	0.983
DistilBERT-base	0.940	0.926	0.926	0.925	0.901
Logistic Regression	0.899	0.898	0.898	0.899	0.864
SVM	0.934	0.933	0.933	0.933	0.911
Naive Bayes	0.860	0.857	0.857	0.857	0.810
Random Forest	0.959	0.959	0.953	0.953	0.945
Gradient Boosting	0.959	0.959	0.958	0.958	0.945
XGBoost	0.968	0.968	0.968	0.968	0.957
LightGBM	0.969	0.969	0.969	0.969	0.958
MLP	0.964	0.964	0.963	0.963	0.952
KNN	0.799	0.798	0.798	0.796	0.730

Table A.7

Skill Level Performance of Fine-Tuned Small Language Models for PSAT

Model	Skill 1	Skill 2	Skill 3	Skill 4	Skill 5	Skill 6	Skill 7	Skill 8	Skill 9	Skill 10
BERT-base	0.996	0.992	0.997	0.692	0.250	0.991	1.000	0.737	0.924	0.986

BERT-large	0.992	0.992	0.981	0.200	0.075	1.000	0.996	0.630	0.672	1.000
ALBERT-base	0.988	0.976	0.968	0.824	0.797	0.995	1.000	0.993	0.981	1.000
ConvBERT	0.992	0.996	0.972	0.150	0.678	1.000	1.000	0.741	0.900	1.000
All-DistilRoBERTa	0.992	0.992	0.963	0.108	0.683	0.926	1.000	0.937	0.917	0.986
ELECTRA-base	0.988	0.992	0.997	1.000	0.974	1.000	1.000	0.993	0.987	1.000
ELECTRA-small	0.988	0.988	0.672	0.000	0.000	1.000	1.000	0.619	0.653	0.839
RoBERTa-base	0.992	0.992	0.955	0.333	0.774	1.000	1.000	0.997	0.993	1.000
RoBERTa-large	0.996	0.996	0.991	0.986	0.974	1.000	1.000	0.997	1.000	1.000
DeBERTa-base	0.996	0.996	0.984	0.867	0.900	0.995	1.000	0.984	0.980	1.000
DeBERTa-large	0.992	0.984	1.000	0.056	0.798	0.995	1.000	0.800	0.695	1.000
DistilBERT-base	0.996	0.996	0.991	0.824	0.424	0.995	0.996	0.904	0.746	1.000

Note. Skill 1 = Boundaries; Skill 2 = Form, Structure, and Sense; Skill 3 = Command of Evidence; Skill 4 = Inferences; Skill 5 = Central Ideas and Details; Skill 6 = Transitions; Skill 7 = Rhetorical Synthesis; Skill 8 = Words in Context; Skill 9 = Text Structure and Purpose; Skill 10 = Cross-Text Connections.

Figure A.1

PCA Projection of Embeddings for Skill 4 (Inferences) vs. Skill 8 (Words in Context) for SAT and PSAT Items

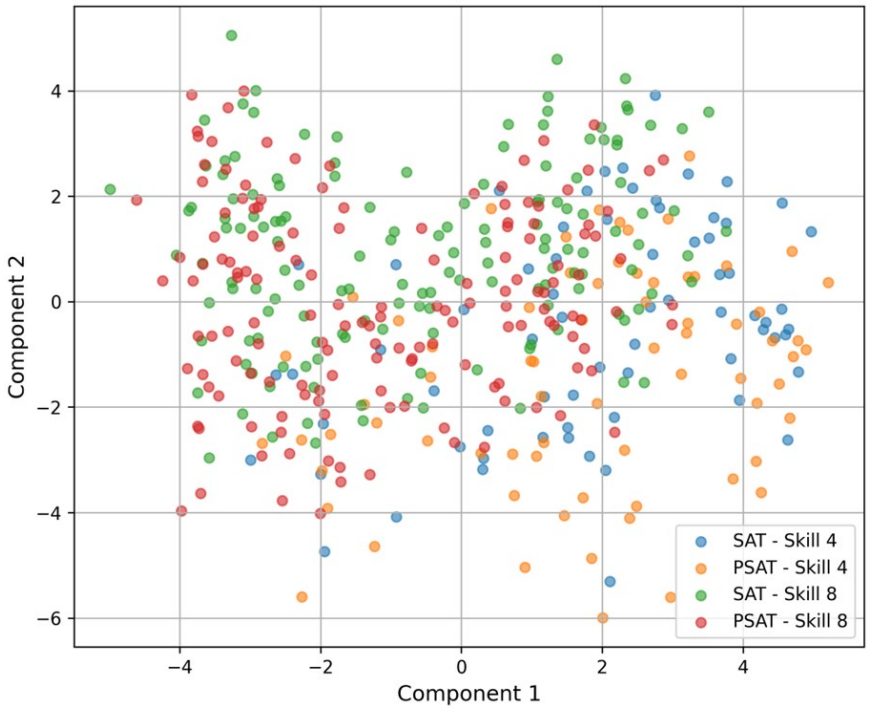

Figure A.2

t-SNE Projection of Embeddings for Skill 4 (Inferences) vs. Skill 8 (Words in Context) for SAT and PSAT Items

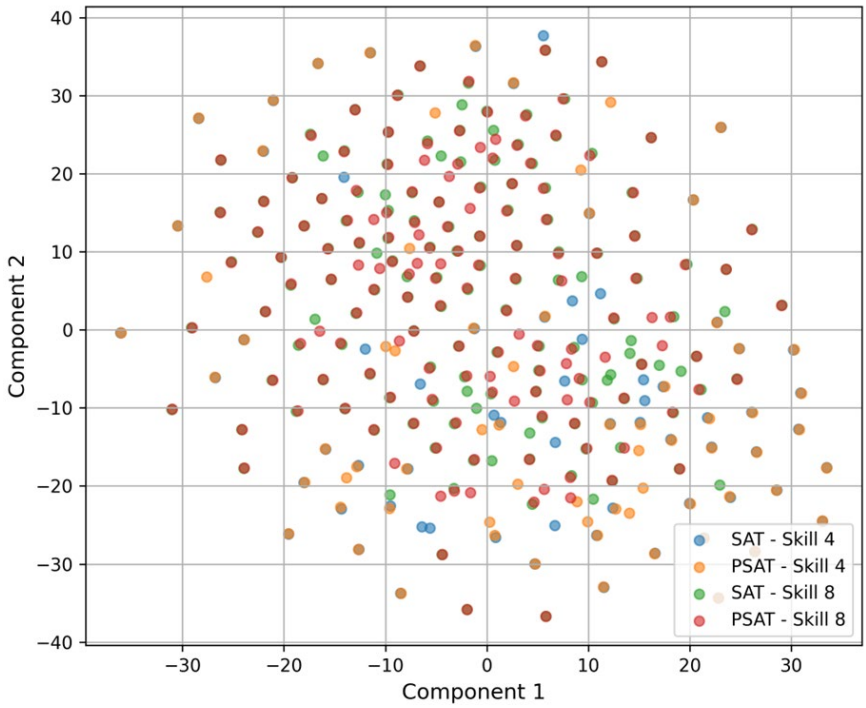

Figure A.3

ISOMAP Projection of Embeddings for Skill 4 (Inferences) vs. Skill 8 (Words in Context) for SAT and PSAT Items

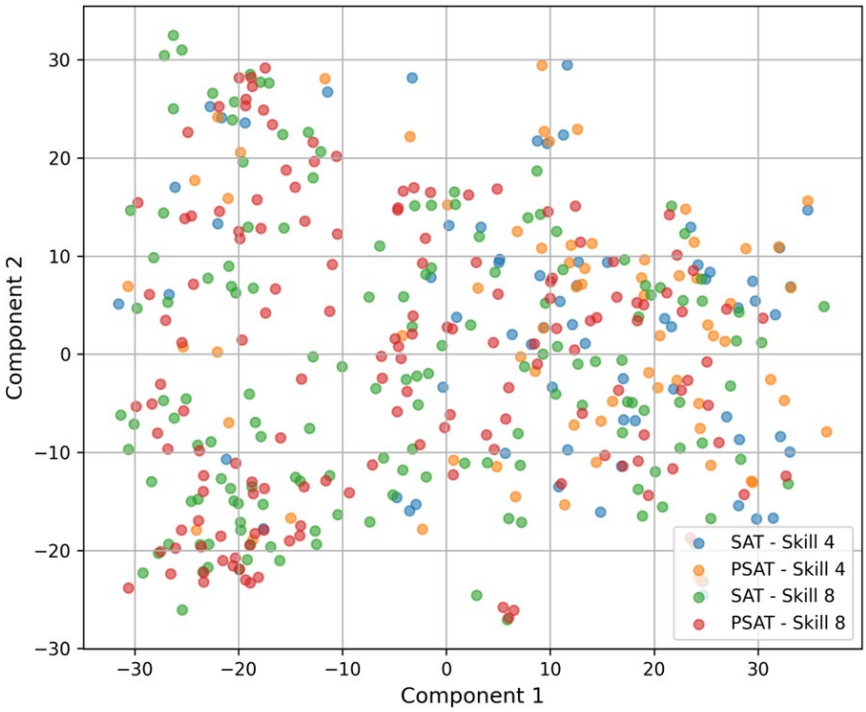

Figure A.4

PCA Projection of Embeddings for Skill 5 (Central Ideas and Details) vs. Skill 8(Words in Context) for SAT and PSAT Items

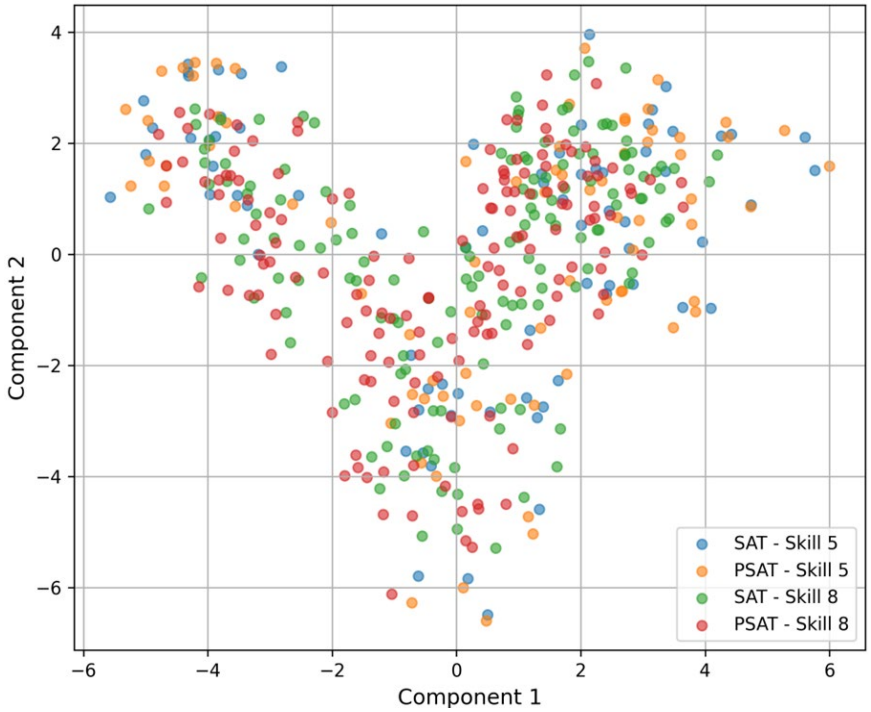

Figure A.5

t-SNE Projection of Embeddings for Skill 5 (Central Ideas and Details) vs. Skill 8 (Words in Context) for SAT and PSAT Items

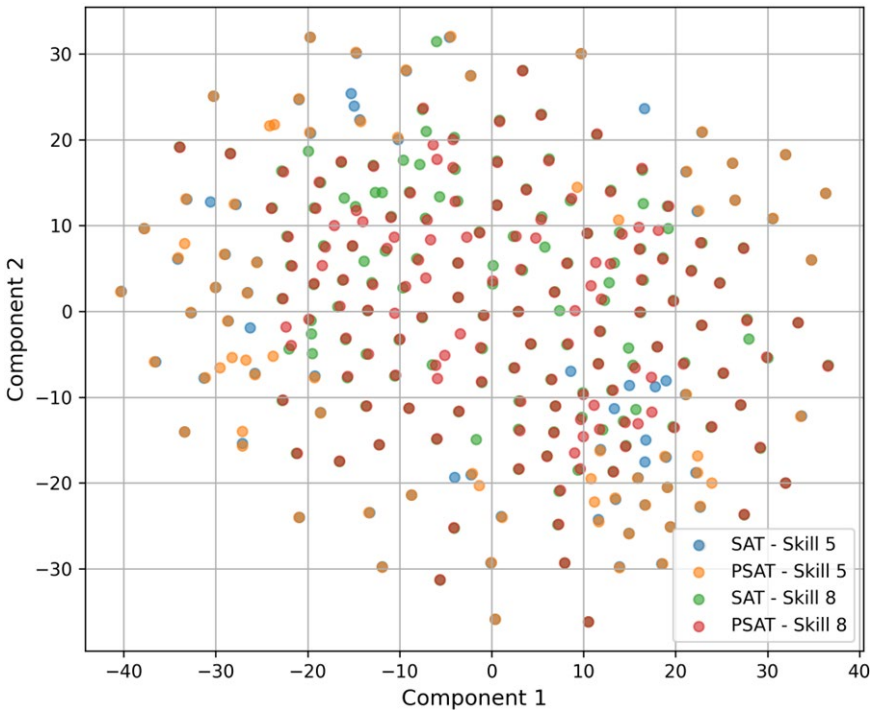

Figure A.6

ISOMAP Projection of Embeddings for Skill 5 (Central Ideas and Details) vs. Skill 8 (Words in Context) for SAT and PSAT Items

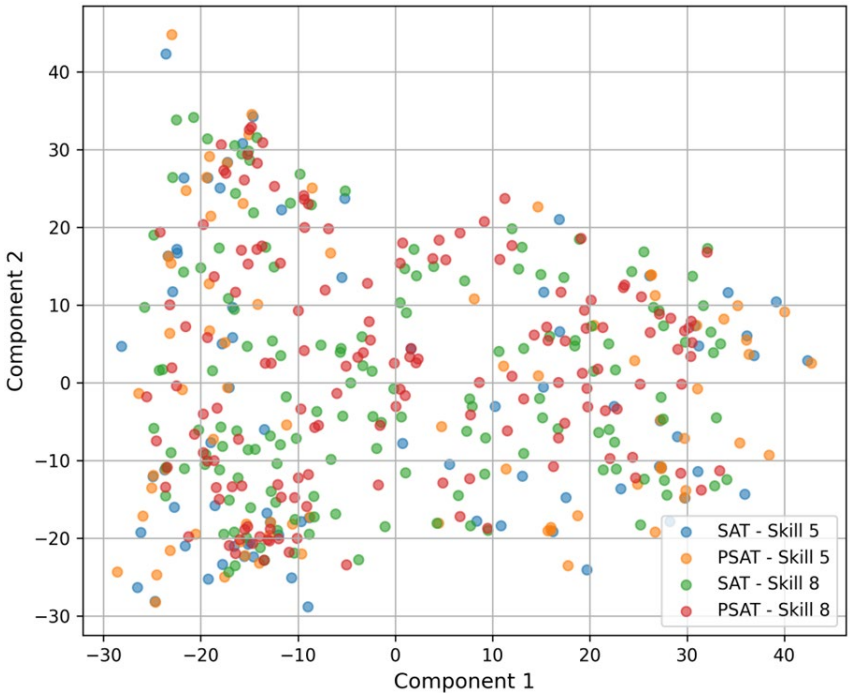

Table A.8

KL Divergence betweeen PSAT Skill 4 and Each SAT Skill

From	To	KL divergence
PSAT skill 4	SAT skill 1	32.927
PSAT skill 4	SAT skill 2	38.059
PSAT skill 4	SAT skill 4	42.588
PSAT skill 4	SAT skill 4	44.503
PSAT skill 4	SAT skill 5	40.996
PSAT skill 4	SAT skill 6	13.610
PSAT skill 4	SAT skill 7	26.869
PSAT skill 4	SAT skill 8	17.986
PSAT skill 4	SAT skill 9	44.342
PSAT skill 4	SAT skill 10	74.312

Table A.9

KL Divergence betweeen PSAT Skill 5 and Each SAT Skill

From	To	KL divergence
PSAT skill 5	SAT skill 1	44.096
PSAT skill 5	SAT skill 2	48.358
PSAT skill 5	SAT skill 3	48.800
PSAT skill 5	SAT skill 4	65.873
PSAT skill 5	SAT skill 5	41.134
PSAT skill 5	SAT skill 6	44.554
PSAT skill 5	SAT skill 7	40.371
PSAT skill 5	SAT skill 8	25.491
PSAT skill 5	SAT skill 9	43.649
PSAT skill 5	SAT skill 10	83.533