

Auxiliary Information for Semantic Clustering of Assessment Items

Josiah Hunsberger¹ Aquia Richburg² Macus Walker²

¹James Madison University, ²National Commission on Certification of Physician Assistants

Abstract

This study identifies a selected transformer-based clustering model for operational medical assessment items that balances within blueprint-topic proximity with cluster separation. Operational analyses showed supplementary alignment with blueprint structure, flagged isolated content areas that may need additional item coverage, and identified dispersed topics for subject matter expert (SME) review.

1 Introduction

Operational assessment programs depend on item banks that are large enough to support repeated form assembly, secure enough to protect score interpretations, and organized enough to maintain content coverage. In high-stakes testing contexts, this organization is usually supported through blueprint classifications such as domains, topics, and cognitive task categories. These classifications are essential because they define the content to be sampled, guide form assembly, and support validity arguments about the intended interpretation and use of scores (Downing, 2006; Haladyna, 2015). However, blueprint categories may not fully represent the finer-grained relationships among items in an operational item bank.

This limitation is especially relevant when items are treated as belonging primarily to a single category for reporting or assembly, even though the underlying content may overlap across clinical scenarios, reasoning tasks, or intended constructs. For example, items assigned to the same broad domain may target substantially different topics, whereas items assigned to different blueprint locations may share meaningful clinical or conceptual structure. In such cases, blueprint classifications remain necessary, but they may not provide enough information to identify substitute items, underdeveloped content areas, overlapping item groups, or item

clusters that could support recommendation and item-bank maintenance.

Embedding-based methods provide one possible way to supplement traditional item-bank organization. Transformer-based language models convert item text into contextual numeric representations that can be used to model semantic similarity among items (Devlin et al., 2019; Vaswani et al., 2017). These representations have been used in assessment contexts for tasks such as identifying overlapping or enemy items and supporting content alignment (Becker & Kao, 2022; Butterfuss & Doran, 2024; Fang, 2020; Mao, 2021). However, item stems are rarely the only source of information available in operational testing programs. Items are typically accompanied by auxiliary information, including blueprint labels, cognitive task classifications, psychometric variables, and sometimes explanatory rationales. These sources may contain information about the item that is not fully captured by stem text alone.

The present study evaluates whether auxiliary item information improves semantic item clustering for operational assessment use. This study asks whether embedding-based clustering can provide additional evidence about item-bank structure. Specifically, we compare a stem-only clustering baseline with clustering models that incorporate item rationales, blueprint topic, cognitive task, and standardized item difficulty. We then apply the best model to operational use cases involving blueprint alignment and item-bank health.

The study addresses two research questions:

1. To what extent does incorporating auxiliary item information improve the quality and stability of semantic item clusters compared with item-text-only clustering?
2. How does the selected cluster structure align with the operational blueprint, and what does it suggest about item-bank health, content

coverage, and potential recommendation use cases?

2 Method

2.1 Item corpus and auxiliary information

The item corpus consisted of operational multiple-choice medical assessment items available for administration between 2023 and 2025 as part of a longitudinal certification examination administered across repeated testing occasions. Each item included a stem, five response options, a keyed response, and an item rationale. The item bank spanned 15 content domains and 235 blueprint topics. In addition to text, each item was associated with several auxiliary variables used in operational assessment processes: blueprint topic, cognitive task category, and standardized item difficulty. These variables represented the intended content target, the cognitive or clinical action evaluated by the item, and psychometric difficulty, respectively.

Rationales were included and evaluated as a non-traditional source of auxiliary information. Rationales are explanatory texts provided after examinees respond and describe the item’s purpose, keyed response, and distractors. Because rationales may identify the underlying concept more explicitly than the item stem, they were included as a supplemental item representation. Citations and references were removed before embedding.

2.2 Embedding, dimensionality reduction, and clustering

Item stems and rationales were transformed into embeddings using the BioClinical ModernBERT (Sounack et al., 2025) model. Because high-dimensional embedding spaces can make item-to-item distances less informative and more difficult to cluster, dimensionality reduction was applied before clustering. Stem and rationale embeddings were reduced using Uniform Manifold Approximation and Projection (UMAP), which is commonly used to reduce high-dimensional embeddings while preserving local neighborhood structure (McInnes et al., 2020). For the stem and rationale embeddings, UMAP was applied with 25 and 10 output dimensions, respectively, using $n_neighbors = 3$, $min_dist = 0$, and the correlation metric.

Clustering was conducted with HDBSCAN, a density-based clustering algorithm that can identify irregularly shaped clusters and classify items as noise (Campello et al., 2015). Identifying noise

was important because some items may not belong to a sufficiently dense semantic group, indicating potential isolation from other content in the item bank. HDBSCAN was implemented with $min_cluster_size = 7$, $min_samples = 2$, and $cluster_selection_method = "eom"$. The $min_cluster_size$ hyperparameter controls the smallest allowable cluster and thus strongly influences granularity.

For each clustering condition, HDBSCAN was fit using a precomputed item-to-item distance matrix. Because the model conditions combined text embeddings with auxiliary item information measured on different scales, all inputs were first converted into pairwise distance matrices. Separate distance matrices were computed for stem embeddings and rationale embeddings. For embedding-based inputs, item-to-item distance was computed as

$$Avg(x, y) = \left(\frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2 \right)^{1/2}, \quad (1)$$

where x and y are item embedding vectors and n is the number of embedding dimensions. Each distance matrix was rescaled to the 0–1 range by dividing all pairwise distances by the maximum observed distance in that matrix.

Following similar methodology from Stadlthanner et al. (2024), we incorporated auxiliary item features as mixed data types, including categorical variables such as topic and task and continuous variables such as standardized item difficulty. Therefore, Gower distance was used to construct auxiliary distance matrices (Gower, 1971):

$$D_{Auxiliary}(x, y) = 1 - \frac{1}{n} \sum_{i=1}^n s_i(x_i, y_i), \quad (2)$$

where $s_i(x_i, y_i)$ is the feature-level similarity between items. Feature-level similarity was defined according to feature type. For categorical variables, two items were treated as perfectly similar if they shared the same category and completely dissimilar if they differed. For numeric variables, similarity decreased as the absolute difference between the two item values increased, after scaling that difference by the observed range of the feature.

For each clustering condition, the final distance matrix was formed as a weighted combination of

the available component matrices:

$$\begin{aligned} \mathbf{W} = & [1 - (w_R j + w_A k)] \mathbf{D}_S \\ & + (w_R j) \mathbf{D}_R \\ & + (w_A k) \mathbf{D}_A, \end{aligned} \quad (3)$$

where $\mathbf{D}_S$, $\mathbf{D}_R$, and $\mathbf{D}_A$ represent the stem, rationale, and auxiliary distance matrices, respectively. The weights ($w_R = .10$) and ($w_A = .10$) were assigned to the rationale matrix and to each auxiliary-information source. The indicator j denotes whether rationale embeddings were included, and k reflects the number of auxiliary-information sources included in the clustering condition. The remaining weight was assigned to the stem matrix. Across conditions, the stem weight ranged from 1.00 in the stem-only baseline model to 0.60 when rationale embeddings and all three auxiliary sources were included. In total, 16 clustering conditions were evaluated, including the stem-only baseline and all combinations of rationale, topic, task, and standardized difficulty added to the stem representation.

2.3 Part 1: Model Evaluation

Part 1 used a subsampling-based model selection design. We generated 1,000 independent subsamples of the item bank. Each subsample selected 160 topics without replacement, representing approximately 68% of available topics. Each of the 16 clustering conditions was applied to each subsample.

Two primary validation metrics were computed for each model condition and subsample. Silhouette evaluates whether items are closer to other items in their assigned cluster than to items in other clusters, with higher values indicating stronger cluster separation (Rousseeuw, 1987). Silhouette score was computed as:

$$\text{Silhouette}(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}, \quad (4)$$

where, for data point x in $\mathbf{W}$, $a(x)$ is the average of all the distances between point x and all other points in its assigned cluster. $b(x)$ is the smallest average distance between x and the points in any other cluster.

Second, we propose an item-family distance metric, referred to as FamDist. In this testing context, item families were operationally represented by blueprint topics. For each family F containing n

items, mean pairwise distance was computed as:

$$\text{FamDist}(F) = \frac{1}{\binom{n}{2}} \sum_{i < j} \mathbf{W}(x_i, x_j), \quad (5)$$

where $\mathbf{W}(x_i, x_j)$ is the distance between items x_i and x_j . Lower FamDist values indicate that items within the same blueprint family are closer in the model space, regardless of whether they were assigned to clusters. Together, these metrics captured complementary forms of evidence: Silhouette served as an internal metric of cluster separation, whereas FamDist served as an external metric of whether operationally associated items were grouped closely in the model space.

The championed model was selected by considering both mean performance and stability across subsamples. A desirable model would reduce FamDist relative to the stem-only baseline while preserving acceptable cluster separation. Because these two criteria represent different forms of evidence, the championed model was not expected to maximize all outcomes simultaneously. Instead, model selection focused on identifying a defensible operational tradeoff among within-family proximity, cluster separation, and the proportion of items retained in clusters.

2.4 Part 2: Operational Application

The championed model was applied to the full pool of 1,772 items. First, blueprint alignment was evaluated by comparing cluster membership with existing blueprint topic classifications. This analysis evaluated the purity of each blueprint topic, not the purity of the clusters themselves. For each topic, topic purity was defined as n_t^{dom}/n_t , where n_t^{dom} is the number of items for topic t assigned to the dominant cluster for that topic, and n_t is the total number of items for topic t . Topic purity and FamDist were used jointly to evaluate blueprint alignment, with purity greater than .50 and low FamDist indicating stronger alignment.

Low-purity topics consisted of topics in which fewer than half of the items were assigned to a single dominant cluster, either because the items were distributed across multiple clusters or because many items were not assigned to sufficiently dense clusters. Topics with many noise items and low FamDist were interpreted as potential coherent but underdeveloped content areas with few plausible substitutes for form assembly or recommendation. Low-purity topics with high FamDist were flagged

Table 1: Cluster Metric Values

Model	FamDist	Silhouette
Stem	0.42 ± .001	0.73 ± .001
Stem + Task + Std_B	0.35 ± .001	0.68 ± .001
Stem + Topic + Task + Std_B	0.31 ± .001	0.59 ± .001
Stem + Rationale + Topic + Task + Std_B	0.30 ± .001	0.56 ± .001

Note. Values are means ± confidence intervals. Lower values are preferable for FamDist, whereas higher values are preferable for Silhouette.

for SME review because they may reflect heterogeneous content or inconsistent representation within the same blueprint topic.

FamDist cutoffs were computed separately for the blueprint-alignment and item-bank-health analyses because the purity threshold defined distinct subsets of blueprint topics. Although topic purity and FamDist were expected to be related, they were not redundant: topics concentrated within a dominant cluster may still vary in within-topic distance. Computing FamDist cutoffs within each subset allowed low and high FamDist values to be interpreted relative to comparable topics rather than applying a single threshold across different groups.

Finally, de-identified examinee response data were linked to the final cluster assignments as a preliminary investigation of whether the identified semantic structure corresponded to interpretable response patterns. Responses were ordered according to each examinee's exposure to items within a selected cluster, regardless of the specific item administered. Cumulative performance and classification accuracy were then examined as additional cluster items were administered. Classification accuracy was defined as whether an examinee's classification at a given administration position matched their classification based on their final cumulative score within the cluster. These analyses were intended as preliminary evidence for the feasibility of a cluster-based recommender system rather than evidence that items within a cluster were psychometrically equivalent.

3 Results

3.1 Part 1: Model Evaluation

The model comparisons showed a consistent trade-off between within-family proximity and cluster separation (see Figure 1).

The stem-only baseline produced the highest FamDist, indicating that items from the same

Table 2: Variation in Cluster Metrics

Model	FamDist <i>SD</i>	Silhouette <i>SD</i>
Stem	0.020	0.011
Stem + Task + Std_B	0.015	0.021
Stem + Topic + Task + Std_B	0.013	0.032
Stem + Rationale + Topic + Task + Std_B	0.013	0.026

blueprint family were farther apart in the stem-only model space than in the auxiliary-informed model spaces. Every auxiliary-information model reduced FamDist relative to the stem-only baseline. This pattern suggests that auxiliary information added meaningful structure beyond the stem text alone.

However, reductions in FamDist were generally accompanied by lower Silhouette scores. The stem-only baseline produced the highest average Silhouette, indicating the strongest separation among clusters. Adding auxiliary information generally pulled related items closer together but also reduced the distinctiveness of the cluster structure. This pattern was especially clear for models that included topic information. Topic reduced FamDist substantially, but it also produced larger reductions in Silhouette than task or standardized difficulty.

Among single-source auxiliary models, task preserved cluster separation most effectively while still reducing FamDist (see Table 1). Topic produced the strongest reduction in FamDist among the single-source models but also imposed a larger penalty on Silhouette. Standardized difficulty provided a moderate improvement in FamDist with a relatively smaller loss in Silhouette. Rationale information produced the smallest reduction in FamDist and a meaningful reduction in Silhouette, suggesting that rationale embeddings did not provide enough unique clustering information to justify their inclusion in the final model.

The dispersion results supported the same interpretation. Auxiliary-information models generally reduced the variability of FamDist across subsamples, indicating more stable within-family proximity (see Table 2). FamDist SD decreased from 0.020 for the stem-only baseline to 0.013 for both the Stem + Rationale + Topic + Task + Std_B model and the full auxiliary-information model. In contrast, these models tended to increase the variability of Silhouette, indicating less stable cluster separation across subsamples. Silhouette SD was lowest for the stem-only baseline at 0.011 and high-

est for the Stem + Topic + Task + Std_B model at 0.032. Thus, the best model was not the one with the lowest FamDist or the highest Silhouette in isolation, but the model that provided the most defensible balance between these criteria.

Silhouette and FamDist capture complementary information. Silhouette rewards compact, well-separated clusters, whereas FamDist rewards proximity among items within the same blueprint topic; optimizing either in isolation may come at the expense of the other. Model selection therefore focused on identifying a defensible balance between geometric cluster quality and within-topic proximity. Three candidate championed models were retained for closer comparison: Stem + Task + Std_B, Stem + Topic + Task + Std_B, and the full model including Stem + Rationale + Topic + Task + Std_B. The Stem + Task + Std_B model was the most conservative option. It preserved the highest Silhouette among the three candidates and retained the largest proportion of clustered items, but it produced less improvement in FamDist than the more complex models.

The full model produced the lowest FamDist, but it also produced the lowest Silhouette, suggesting that the additional reduction in within-family distance came at the cost of weaker cluster separation. The full model also provided only modest improvement over the four-source model that excluded rationale information.

The Stem + Topic + Task + Std_B model was selected as the championed model (see Figure 2). This model reduced FamDist more substantially than the Stem + Task + Std_B model while avoiding the additional loss in Silhouette observed for the full model. It also retained a usable proportion of clustered items and produced cluster sizes that were operationally interpretable. The selected model therefore represented a practical compromise: it incorporated blueprint topic, cognitive task, and standardized difficulty information, but excluded rationale embeddings because they did not provide enough incremental benefit.

Part 2: Operational Application

The championed Stem + Topic + Task + Std_B model was applied to the full pool of 1,772 items, identifying 66 clusters containing 1,402 items (79.1%); the remaining 370 items (20.9%) were classified as noise. Cluster sizes ranged from 10 to 64 items ($M = 21.24$), with an overall FamDist of 0.312 and Silhouette of 0.547. Clusters repre-

Table 3: Classification and Cutoff

Classification	Rule	Topics	Mean	Min	Max
<i>Purity</i>					
Moderate	$< .925$	32	.66	.50	.83
High	$\geq .925$	115	1.00	1.00	1.00
<i>FamDist</i>					
Low	$\leq .025$	36	.02	.003	.02
Moderate	$.025 < x < .148$	75	.08	.03	.15
High	$\geq .148$	36	.25	.15	.49

Note. Classification rules were based on empirical quartile cutoffs from the final clustering solution.

sented an average of 3.89 blueprint topics and 1.83 domains, with an average topic purity of 0.779, suggesting that the resulting groupings were related to, but not redundant with, the operational blueprint. The final cluster solution was then evaluated for blueprint alignment. This analysis focused on topics in which at least 50% of items were assigned to clusters. Topic purity and FamDist were interpreted jointly: purity indicated whether items from the same topic were concentrated in a dominant cluster, whereas FamDist indicated whether items from the same topic were close together in the model space.

As shown in Figure 3, most topics demonstrated high purity, including 67 with moderate FamDist (e.g., asthma) and 36 with low FamDist (e.g., rubella). Eight topics showed moderate purity and moderate FamDist (e.g., PTSD), whereas 12 showed high purity but high FamDist (e.g., sickle cell anemia). Overall, most retained blueprint topics were represented coherently by the clustering model, although some may reflect within-topic heterogeneity or broader content definitions.

The item-bank health analysis focused on 88 low-purity topics, defined as topics in which fewer than half of the items were assigned to clusters. Among the 88 low-purity topics, 21 showed low FamDist ($\leq .05$) despite most or all items being classified as noise, suggesting coherent but potentially underdeveloped content areas. In contrast, 21 showed high FamDist ($\geq .20$), suggesting greater within-topic dispersion and potential need for SME review.¹

Examinee response data provided preliminary evidence that performance within the selected semantic cluster became clearer as additional responses accumulated. Figure 4 displays the distri-

¹FamDist cutoffs were computed separately for high-purity topics in the blueprint alignment analysis and low-purity topics in the item-bank health analysis.

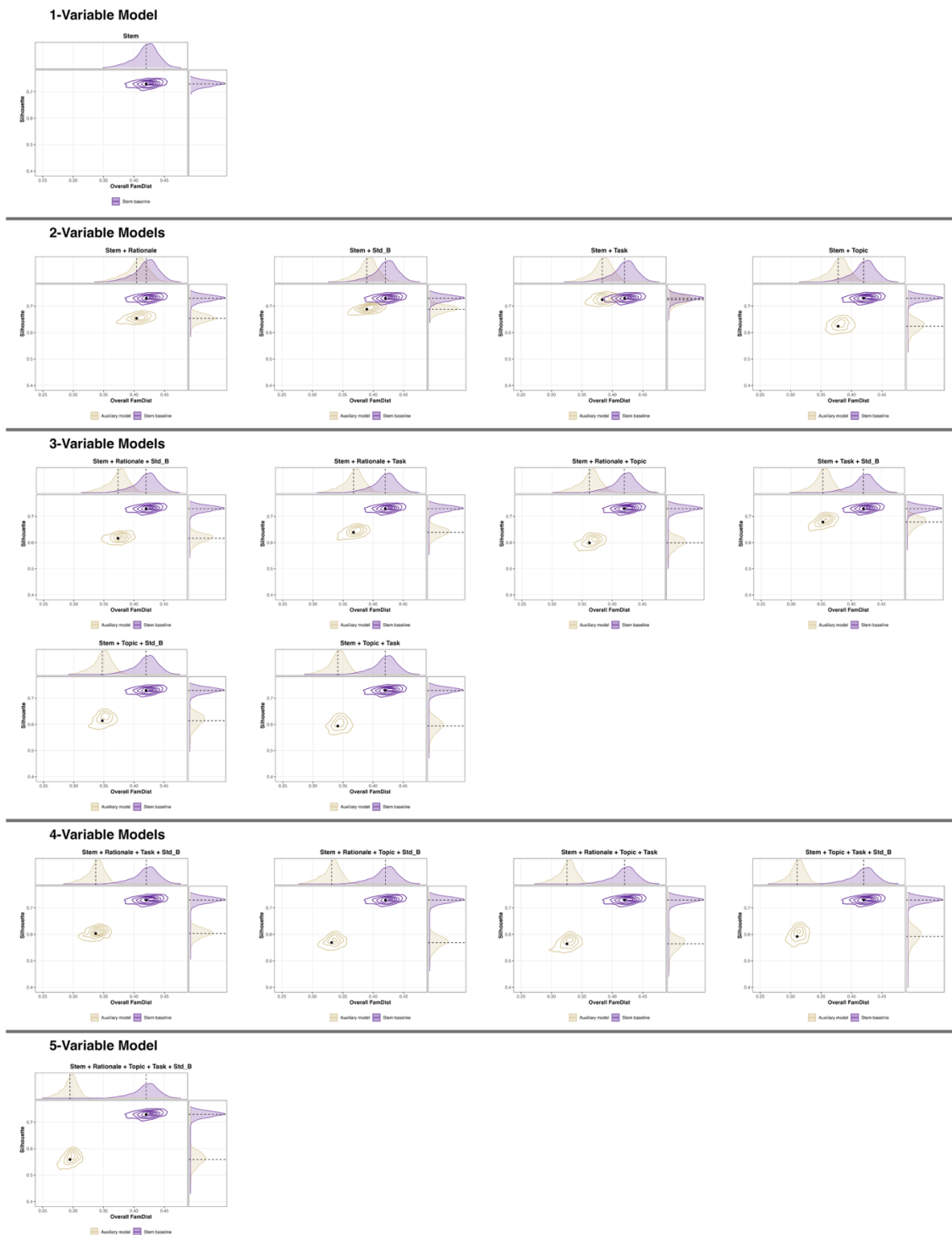


Figure 1: Joint distributions of FamDist and Silhouette by clustering model and number of information sources.

bution of the cumulative number correct across longitudinal administration positions. Across the more densely represented positions, cumulative scores were concentrated toward the upper end of the possible score range.

At the individual-examinee level, cumulative proportion correct was highly constrained during

the earliest administration positions because only a limited number of cumulative scores were mathematically possible. As additional responses accumulated, trajectories became progressively more differentiated. Examinees who ultimately performed at or above 80% generally remained toward the upper portion of the distribution, whereas

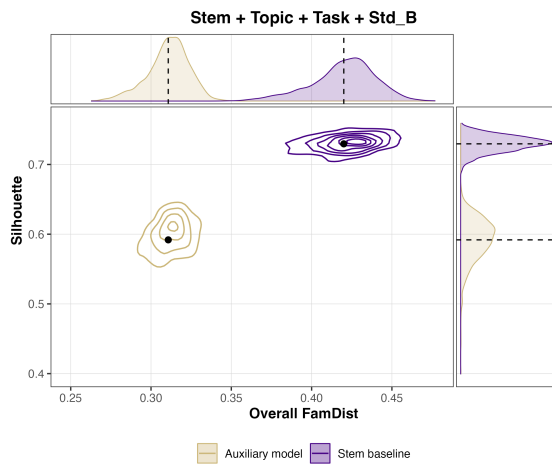


Figure 2: Joint distributions of FamDist and Silhouette of the championed model.

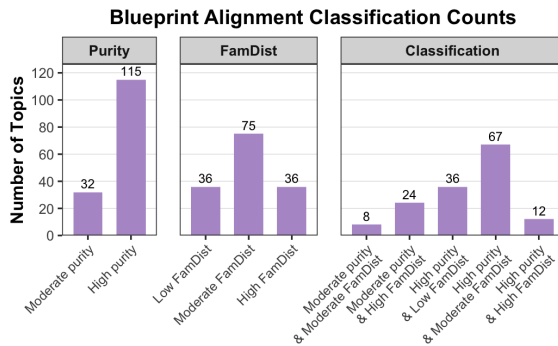


Figure 3: Bar chart of topic classifications.

examinees finishing below 70% increasingly occupied lower trajectories. Repeated early incorrect responses, particularly when cumulative performance declined to approximately 50% or lower within the first three items, appeared more characteristic of examinees who ultimately demonstrated lower within-cluster performance (see Figure 5).

Classification accuracy generally improved as additional responses became available, although the pattern differed across final performance groups. Examinees who ultimately scored at or above 80% were classified with relatively high accuracy throughout most administration positions. Classification of examinees in the 70% to < 80% range was considerably less stable during the earlier positions, while accuracy for examinees ultimately scoring below 70% generally increased as additional responses accumulated (Figure 6). Estimates at later administration positions should be interpreted cautiously because progressively fewer examinees received larger numbers of items from the cluster.

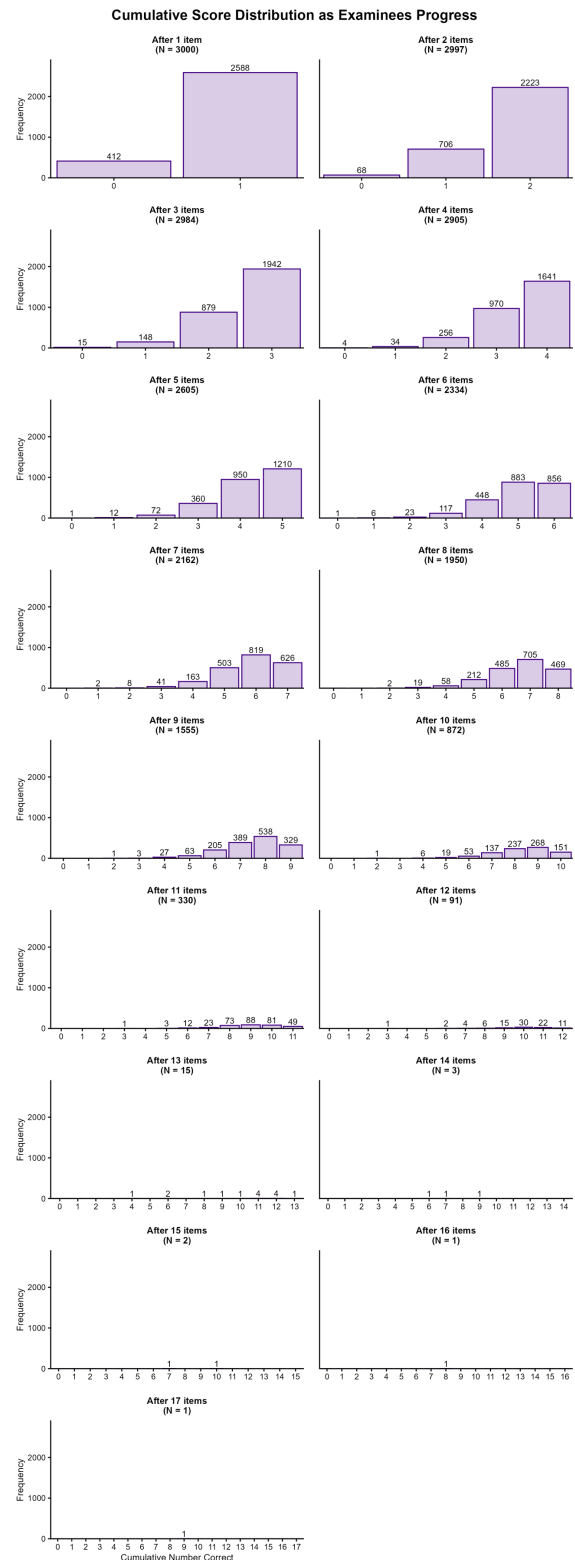


Figure 4: Cumulative score distribution of examinee responses.

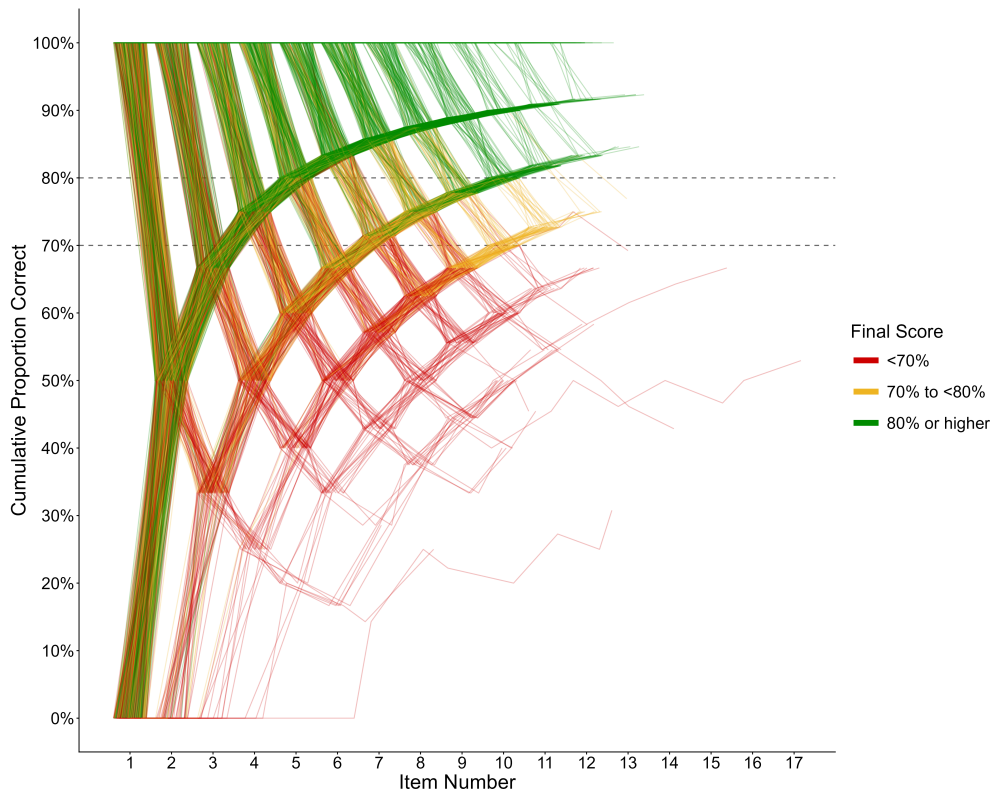


Figure 5: Examinee cumulative score progression.

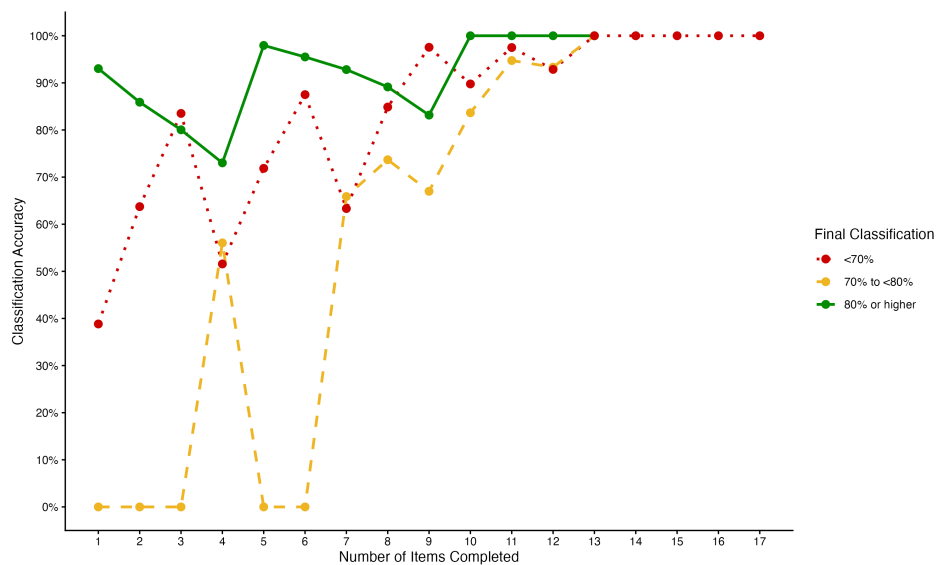


Figure 6: Average classification accuracy across administered items.

4 Discussion and Conclusion

This study evaluated whether auxiliary item information improves semantic clustering for operational assessment use. Auxiliary-informed models consistently reduced FamDist, indicating that items within the same blueprint family, but also generally reduced Silhouette scores. Thus, improving alignment among operationally related items often came

at the cost of weaker cluster separation. The championed model therefore reflects a balance between blueprint alignment and empirical cluster structure.

The results also indicate that not all auxiliary sources contributed equally. Topic information substantially reduced FamDist but also weakened cluster separation, whereas task and standardized difficulty provided a more favorable balance between

the two metrics. Rationales contributed limited unique information beyond item stems and were therefore excluded from the championed model.

The results in Part 2 show how auxiliary-informed clustering moves beyond global model evaluation to operational applications. The identified clusters were related to, but not redundant with, the blueprint. Blueprint alignment identified topics supported by the cluster structure and others that may warrant SME review, while the item-bank health analysis distinguished potentially isolated content groups from topics showing greater within-topic dispersion.

The examinee-response analysis provides an initial indication that the identified clusters may also offer a useful structure for organizing performance information. As additional responses accumulated, cumulative trajectories became more differentiated and classification tended to correspond more closely with examinees' eventual performance within the cluster, suggesting that cluster-level response histories may become increasingly informative as evidence accumulates.

The response trajectories also suggest that the amount of evidence required may not be the same for every examinee. By approximately 10 observations in the selected cluster, cumulative performance was more clearly differentiated for many examinees, but some persistently low-performing trajectories appeared informative earlier. Rather than adopting a single fixed cutoff, future recommendation systems could investigate adaptive classification-based stopping rules similar to those used in computerized adaptive testing.

These findings contribute to NLP-integrated assessment by emphasizing that embedding-based methods should be evaluated using measurement-relevant evidence. In assessment settings, clusters must also be interpretable relative to blueprint structures, psychometric information, intended operational uses, and SME judgement.

The primary limitation is the study's context specificity. The study was conducted in one operational medical assessment context, so the selected weighting scheme and championed model may not generalize directly to other testing programs. Cluster assignments should supplement rather than replace SME judgement.

The examinee-response analysis also has several important limitations. These response patterns should nevertheless be interpreted as preliminary evidence rather than evidence that items within a

cluster are psychometrically interchangeable. Examinees received different items, later administration positions contained substantially smaller samples, and item difficulty remained relevant despite its inclusion in the clustering model. The longitudinal design may also introduce potential learning effects because examinees received rationales after each item. Consequently, changes in cumulative performance cannot be attributed solely to repeated exposure to semantically related content.

The preliminary response analysis therefore begins the validation of clusters for potential examinee-facing applications but does not establish their effectiveness for recommendation. Future work should evaluate whether cluster-level response histories predict subsequent performance on related items while accounting for item difficulty, examinee ability, and longitudinal learning.

Overall, the findings support the use of selected auxiliary information in semantic item clustering. A model incorporating stem embeddings, topic, task, and standardized difficulty provided the strongest operational balance in this study. The resulting clusters offer a practical foundation for blueprint review, item-bank health monitoring, and future validation with response data.

References

- Becker, K. A., & Kao, S. (2022). Identifying Enemy Item Pairs using Natural Language Processing. *Journal of Applied Testing Technology*, 23(Special Issue 1), 53–64.
- Butterfuss, R., & Doran, H. (2024). An Application of Text Embeddings to Support Alignment of Educational Content Standards. *Educational Measurement: Issues and Practice*, emip.12641. <https://doi.org/10.1111/emip.12641>
- Campello, R. J. G. B., Moulavi, D., Zimek, A., & Sander, J. (2015). Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. *ACM Transactions on Knowledge Discovery from Data*, 10(1), 1–51. <https://doi.org/10.1145/2733381>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding* (arXiv:1810.04805). arXiv. <http://arxiv.org/abs/1810.04805>
- Downing, S. M. (2006). Twelve Steps for Effective Test Development. In T. M. Haladyna & S. M. Downing (Eds.), *Handbook of test development* (1. publ., repr, pp. 3–25). Routledge.
- Fang, P. (2020). *Automatic enemy item detection using natural language processing*. University of Illinois at

Chicago.

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, 27(4), 857. <https://doi.org/10.2307/2528823>

Haladyna, T. M. (2015). The Importance of Item Development for Validity. In *Developing and validating multiple-choice test items* (Third edition, first issued in paperback). Routledge, Taylor & Francis Group.

Mao, X. (2021). An Exploration of an Integrated Approach for Enemy Item Identification. *Proceedings of the 2021 AERA Annual Meeting*. 2021 AERA Annual Meeting. <https://doi.org/10.3102/1690935>

McInnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* (arXiv:1802.03426). arXiv. <https://doi.org/10.48550/arXiv.1802.03426>

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.

Journal of Computational and Applied Mathematics, 20, 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)

Sounack, T., Davis, J., Durieux, B., Chaffin, A., Pollard, T. J., Lehman, E., Johnson, A. E. W., McDermott, M., Naumann, T., & Lindvall, C. (2025). *BioClinical ModernBERT: A State-of-the-Art Long-Context Encoder for Biomedical and Clinical NLP* (arXiv:2506.10896). arXiv. <https://doi.org/10.48550/arXiv.2506.10896>

Stadlthanner, D., Steinkellner, H., Landschützer, C., & Kaefer (geb. Prims), D. (2024). *A Hierarchical Density-Based Clustering Method Applied to Mixed-Mail in Austria* (4th ed.). Bundesvereinigung Logistik (BVL) e.V. https://doi.org/10.23773/2024_4

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need*. 31, 1–11.