

Pre-Data Embedding Diagnostics for Construct Validity in Multi-Domain Instruments

Youngmi Cho Tong Wu Hyeonjoo Oh

Riverside Insights

✉ youngmi.cho@riversideinsights.com

Abstract

Centroid margin, an embedding-based measure of relative domain distinctiveness, was evaluated as a pre-data item-selection tool for multi-domain instruments. Across four samples, centroid-guided selection improved model fit and discriminant validity, whereas own-domain cosine similarity increased overlap among closely related domains, revealing a key limitation of absolute similarity metrics.

1 Introduction

Developing multi-domain psychological instruments requires two psychometric properties: strong within-domain coherence and clear between-domain distinctiveness. These properties correspond to the core components of construct validity – convergent and discriminant validity – and are especially difficult to optimize jointly when domains are conceptually related. Their evaluation has traditionally depended on item-level statistics derived from pilot data, such as corrected item-total correlations, standardized factor loadings, and differentiation ratios. These diagnostics guide iterative cycles of item revision and additional data collection. Although effective, this process creates a practical bottleneck early in instrument development, when candidate items exist only as text and empirical data are unavailable.

Recent advances in large language model (LLM) embeddings offer a way to conduct part of this evaluation before data collection. The conceptual basis lies in the lexical hypothesis tradition (Galton, 1884; Goldberg, 1993): if language encodes psychologically meaningful variation, geometric relationships among item embeddings in semantic space may anticipate empirical covariation in response data. Earlier approaches

based on latent semantic analysis had limited success in predicting item behavior, particularly in personality domains with abstract and overlapping constructs (Arnulf et al., 2014). Transformer-based models have expanded this capacity by representing contextualized, sentence-level meaning and capturing the psychological content and nuance of items more effectively than earlier distributional methods

Most embedding-based applications in psychometrics rely on cosine similarity, which indexes the absolute proximity of two vectors in semantic space. Cosine similarity has been used to approximate inter-item relationships, recover latent structure, and identify construct redundancy. For example, Wulff & Mata (2025) applied cosine similarity from fine-tuned MPNet embeddings across hundreds of international personality item pool (IPIP) scales to detect jingle-jangle fallacies, showing that embedding-based clustering can meaningfully reduce construct redundancy. Milano et al. (2025) found that cosine similarity matrices derived from RoBERTa embeddings could recover factorial structures through principal component analysis without response data; embedding-based loadings correlating from moderately to strongly with empirical factor loadings across multiple instruments. Feraco and Toffalini (2025) extended this approach in SEMbeddings by fitting confirmatory factor models directly to embedding-based similarity matrices. They found that these matrices approximated empirical inter-item correlations well enough to identify problematic item pairs through modification indices.

At the item level, Voss et al. (2026) found that cosine similarity predicted factor loadings across diverse instruments, particularly when represented as a difference score contrasting within-domain and cross-domain similarity rather than as absolute

within-domain similarity alone. These difference scores were substantially more useful for identifying cross-loading items, suggesting that an item's relative positioning in embedding space provides information that absolute proximity does not. Collectively, this literature indicates that embeddings encode psychologically meaningful structure while also pointing to a limitation that absolute similarity alone may be insufficient for improving discriminant validity in multidomain instruments. This limitation raises a central question:

Which properties of embedding space are most informative for improving construct validity before data collection?

This question is particularly important when construct boundaries are inherently indistinct. The present study examines social-emotional competence (SEC), a construct in which the domains are explicitly conceptualized as fuzzy, prototype-based constructs rather than discrete, mutually exclusive dimensions (Soto et al., 2022). SEC frameworks emerged largely from applied and policy contexts rather than from strong psychometric evidence of construct separability (Kyllonen & Zu, 2025), making discriminant validity more difficult to achieve than in more established domains such as personality. The practical stakes are also substantial: SEC domain scores are widely used to create individual profiles, guide interventions, and monitor change over time (Durlak et al., 2011; Shapiro et al., 2024). These applications assume that domain scores represent meaningfully distinct constructs, creating a need for methods that can sharpen domain boundaries during instrument development.

The present study extends prior work by systematically comparing three embedding-based metrics that capture distinct properties of item representations, moving beyond the dominant focus on cosine similarity. Two metrics are geometric and derived directly from the structure of the embedding space. Own-domain cosine similarity (CS) captures absolute convergence by measuring an item's proximity to the centroid of its intended domain without reference to other domains. Centroid margin (CM) captures relative distinctiveness by quantifying how much closer an item is to its own domain centroid than to the

nearest competing domain centroid, explicitly incorporating cross-domain contrast. The third metric, classifier confidence (CC), is a model-based probabilistic measure. It represents the likelihood that a trained classifier assigns an item to its intended domain based on patterns learned from the full embedding distribution (Debelak et al., 2025). Thus, CS emphasizes within-domain fit, CM emphasizes separation from competing domains, and CC emphasizes the recoverability of domain labels through supervised learning.

To evaluate the three metrics in a realistic development context, we applied them to a fixed item pool from an established SEC instrument. Each metric was used to construct an alternative short form while retaining the item-selection criteria used to develop the instrument's published short form. We then compared the three resulting forms with the published form, which served as the benchmark, across four samples. This design allowed us to evaluate the forms' structural properties and the consistency of the results across samples.

Three research questions guided the analysis. First, to what extent do the three metrics identify the same items as suitable for inclusion in the short form? Second, do embedding-guided selections improve construct validity, as indicated in model fit, factor structure, and discriminant relationships across samples? Third, do embedding-guided selections improve the prediction of real-world outcomes, indicating that embedding-based diagnostics capture practically relevant as well as structural signals?

These questions examine which properties of embedding space are most relevant to construct validity and which diagnostics may be useful before data collection. The diffuse boundaries among SEC domains make this a demanding test of whether embedding-derived information can improve domain distinctiveness.

2 Methods

2.1 Instrument and Benchmark

The Behavioral, Emotional, and Social Skills Inventory (BESSI; Soto et al., 2022) was selected as a practical test case rather than as a target of psychometric critique. BESSI measures five

broad SEC domains (SM: Self-Management, SE: Social Engagement, CO: Cooperation, ER: Emotional Resilience, IN: Innovation) across 32 facets. These constructs are explicitly characterized as fuzzy classes with substantial conceptual overlap (Soto et al., 2022). Its 192-item pool is large enough to support the construction of alternative forms while preserving the original facet structure. In addition, the rigorously validated 45-item short form (BESSI-45; Sewell et al., 2025) provides a strong empirical benchmark for evaluating embedding-guided alternatives.

An ideal evaluation would use a newly developed item pool spanning a wide range of item quality, but such datasets are rarely publicly available. BESSI’s well-documented psychometric properties and publicly available data therefore make it a suitable alternative.

2.2 Data

All data were obtained from the Open Science Framework (OSF¹). The analysis included four samples: high school ($N=499$), college ($N=322$), adult Internet C ($N=400$), and adult Internet D ($N=600$), for a total $N=1,821$. The high school sample also included external criteria for predictive validity analyses: school-reported winter GPA ($N=469$), school-reported spring GPA ($N=458$), and self-reported academic engagement ($N=376$).

2.3 Embedding Model

Text embeddings were computed with the *all-roberta-large-v1* model (Reimers & Gurevych, 2019) from the Sentence-Transformers library. Based on the RoBERTa-large architecture (Liu et al., 2019), the model produces 1,024-dimensional dense vectors optimized for semantic similarity through contrastive training. To align the embeddings with the intended construct meaning and the BESSI response format, each item was embedded within the capacity prompt, “How well can you [item text]?” A total of 150 single-loaded items (i.e., items without substantial cross-loadings in the published factor solution) were encoded in a single pass. The resulting embedding matrix was used in all subsequent analyses.

2.4 Centroid margin (CM)

Centroid margin represents relative semantic distinctiveness: the extent to which an item is positioned within its intended domain relative to competing domains. For each item, we calculated the centroids of every domain as the arithmetic mean of item embeddings. The target item was excluded when calculating the centroid of its assigned domain. The centroid margin for item i was defined as:

$$CM_i = \cos(\mathbf{e}_i, \mathbf{c}_k) - \max_{j \neq k} \cos(\mathbf{e}_i, \mathbf{c}_j),$$

where i indexes the item being evaluated, k denotes its assigned domain, and j denotes a competing domain ($j \neq k$), $\mathbf{e}_i$ is the item’s embedding, $\mathbf{c}_k$ is the centroid of the assigned domain, computed with item i excluded (leave-one-out), and $\mathbf{c}_j$ is the centroid of competing domain j . A positive value indicates that the item is closer to its intended domain than to any alternative. Larger values indicate greater distinctiveness, whereas low or negative values indicate ambiguous cross-domain positioning.

2.5 Classifier confidence (CC)

Classifier confidence represents probabilistic separability: the degree to which a supervised model can distinguish an item’s intended domain from competing domains. An L2-regularized linear classifier was trained on the embeddings of the 150 items, with domain membership as the outcome. The classifier’s decision-function scores were transformed using the softmax function to obtain normalized confidence scores across the five domains. Five-fold cross-validation with balanced class weights was used to generate out-of-fold scores and reduce overfitting and optimistic bias. For each item, the softmax-transformed score assigned to its intended domain in the held-out fold served as the classifier-confidence metric. Higher values indicate clearer separation of the intended domain from competing domains in the classifier’s decision space.

2.6 Own-Domain Cosine Similarity (CS)

Own-domain cosine similarity represents semantic convergence by measuring how closely an item aligns with the centroid of its intended domain. As

¹ <https://osf.io/4zgry/files/osfstorage>

in the centroid-margin calculation, the target item was excluded when calculating its intended-domain centroid. The cosine similarity for item i was defined as:

$$CS_i = \cos(\mathbf{e}_i, \mathbf{c}_k),$$

where $\mathbf{e}_i$ is the item embedding and $\mathbf{c}_k$ is the centroid of its assigned domain. Higher values indicate stronger semantic alignment with the intended domain. Unlike centroid margin, CS does not incorporate competing domains and therefore captures convergence rather than discriminant positioning.

2.7 Embedding-based Selection Conditions

Three alternative 45-item forms were constructed from the BESSI-192 candidate pool under three conditions. In every condition, the four interstitial items from the BESSI-45 benchmark were retained. The remaining 41 items were selected from the 150 single-loaded items while maintaining facet-level requirement.

- C1 (Centroid-guided): Within each facet, items were ranked by centroid margin, and the highest-scoring item(s) were selected. Classifier confidence served as the tiebreaker.
- C2 (Classifier-guided): Within each facet, items assigned to their intended domain were ranked by classifier confidence. If too few items met this criterion, the classification constraint was removed and all items were ranked by classifier confidence.
- C3 (Cosine-guided): Within each facet, items with a positive centroid margin were ranked by own-domain cosine similarity. If too few items met this condition, the margin constraint was removed and all items were ranked by cosine similarity. The positive-margin requirement served only as a floor, excluding items closer to a competing domain than to their intended domain; among eligible items, selection was based solely on cosine similarity.

2.8 Evaluation

Construct validity was evaluated through confirmatory factor analysis (CFA) in *lavaan* (Rosseel, 2012). We used diagonally weighted least squares (DWLS) estimation with ordered

indicators and standardized latent variance (*std.lv* = true), consistent with the benchmark specification. Correlated residuals were included within two-item facets.

Model fit was assessed with the comparative fit index (CFI), Tucker-Lewis index (TLI), root mean square error of approximation (RMSEA), and standardized factor loadings and average variance extracted (AVE). Discriminant validity was indexed by the mean latent factor intercorrelations (Mean IC) and the heterotrait-monotrait ratio (HTMT). Reliability was evaluated with McDonald's ω and Cronbach's α . Drawing on commonly used fit-index change criteria (Cheung & Rensvold, 2002; Chen, 2007), $\Delta CFI > +.010$ and $\Delta RMSEA < -.005$ were interpreted as indicating meaningful improvement in model fit.

Predictive validity was evaluated in the high school sample using adjusted R^2 from regression models predicting winter GPA, spring GPA, and academic engagement. Each domain score was the unweighted mean of its nine items. Gender and grade level were included as covariates in all models.

3 Results

3.1 Metric Convergence

We first examined whether the three metrics converged in their selection outcomes and the characteristics they appear to favor. Across the 150 single-domain candidate items, centroid margins ranged from $-.171$ to $+.312$ ($M = .100$, $SD = .088$); 16% of items had negative margins, indicating cross-domain ambiguity. Classifier accuracy was high overall but lower for SE and CO, the domain pair with the highest factor intercorrelation in the benchmark ($r = .70$; Figure 1). This domain-level correspondence suggests that the embedding diagnostics detected some of the construct overlap observed in response data. Relative to the benchmark, C1 replaced 25 items, C2 replaced 28, and C3 replaced 23, indicating that the metrics prioritized different subsets of the candidate pool.

The metrics showed limited convergence at item-selection level. C1 and C2 shared 51% of items, C1 and C3 shared 49%, and C2 and C3 shared 24%. Only 20% (8 of 41) of the selected items were common to all three conditions.

Benchmark				C1-Centroid-guided Ranked by CM			C2-Classifier-guided Ranked by CC			C3-Cosine-guided Ranked by CS		
Domain	CM	CC	CS	CM	CC	CS	CM	CC	CS	CM	CC	CS
SM	.151	.567	.664	.205	.679	.692	.190	.715	.677	.161	.609	.749
SE	.066	.520	.684	.099	.563	.655	.048	.596	.604	.059	.474	.723
CO	.077	.483	.729	.102	.471	.748	.066	.565	.685	.089	.467	.755
ER	.140	.565	.692	.182	.578	.696	.141	.633	.670	.159	.553	.727
IN	.114	.621	.668	.174	.652	.681	.170	.659	.674	.145	.557	.709
Overall	.111	.552	.687	.154	.591	.694	.125	.636	.663	.124	.534	.733

Table 1: Mean Embedding Metrics by Domain and Condition
Note. CM = Centroid margin; CC = Classifier confidence; CS = Cosine similarity.

Table 1 shows the embedding profiles produced by the three conditions. As intended, each condition achieved the highest overall value on its own target metric (C1: CM = .154; C2: CC = .636; C3: CS = .733). The conditions differed, however, in their effects on the other two metrics. C1 increased centroid margin while also maintaining or improving both classifier confidence and cosine similarity, indicating that gains in construct distinctiveness were achieved without trade-offs in the other embedding properties. C2 increased classifier confidence but reduced cosine similarity, suggesting that greater classification certainty came at the expense of within-domain semantic coherence. C3 produced the highest cosine similarity but yielded only a marginal increase in centroid margin and the lowest classifier confidence. Thus, conditions that optimized a single metric in isolation often traded off against other desirable properties, whereas centroid-guided selection improved construct separability without such costs.

These differences were most apparent for the highly overlapping SE and CO domains (Figure 1). Compared with the benchmark, C1 reduced the SE-CO intercorrelation (.640 vs. .704), whereas C3 increased it (.767 vs. .704). The opposite effects on the SE-CO correlation demonstrate that the selection metric influenced the extent to which these closely related domains remained distinct.

3.2 Construct Validity

Tables 2 and 3 summarize model fit, discriminant validity, convergent validity, and reliability using median values across the four samples. Additional fit indices (TLI, SRMR) are omitted from Table 2 for brevity. Each delta (Δ) represents the value from an embedding-guided condition minus the corresponding benchmark value.

	CFI	RMSEA	Mean IC	HTMT
Benchmark	.865	.071	.521	.508
C1	.887	.064	.448	.446
C2	.846	.066	.478	.473
C3	.866	.068	.498	.493
$\Delta C1$	.022	-.007	-.073	-.062
$\Delta C2$	-.019	-.005	-.043	-.035
$\Delta C3$	.001	-.003	-.023	-.015

Table 2: Median Model Fit and Discriminant Validity by Condition

Condition	Mean Loading	AVE	ω	α
Benchmark	.595	.383	.845	.841
C1	.609	.393	.853	.844
C2	.578	.356	.833	.825
C3	.594	.384	.847	.839
$\Delta C1$	.014	.001	.008	.003
$\Delta C2$	-.017	-.027	-.012	-.016
$\Delta C3$	.001	.001	.002	-.002

Table 3: Median Convergent Validity and Reliability by Condition

C1 produced the most favorable overall profile. Relative to the benchmark, model fit improved on both CFI ($\Delta = +.022$) and RMSEA ($\Delta = -.007$). Discriminant validity also showed the largest gains of any condition, with reductions in both Mean IC ($\Delta \text{Mean IC} = -.073$) and HTMT ($\Delta \text{HTMT} = -.062$). At the same time, convergent validity remained largely unchanged ($\Delta \text{Load} = +.014$, $\Delta \text{AVE} = +.010$), with most of the improvement concentrated in the IN domain, as shown in the standardized factor loadings (diagonal values) in Figure 1. Reliability also remained essentially unchanged ($\omega = .853$ vs. $.845$). Together, these results suggest that C1 improved separation among domains without sacrificing within-domain coherence.

C2 produced a more mixed pattern. RMSEA improved slightly ($\Delta = -.005$), but CFI fell below the benchmark ($\Delta \text{CFI} = -.019$), providing little evidence of an overall improvement in model fit. Discriminant validity improved ($\Delta \text{Mean IC} = -.043$, $\Delta \text{HTMT} = -.035$), though less consistently across domain pairs, and these gains were accompanied by declines in convergent validity ($\Delta \text{Load} = -.017$, $\Delta \text{AVE} = -.027$). Reliability also decreased slightly. Overall, gains in discriminant validity were accompanied by reduction in convergent validity. A stability analysis confirmed that the classifier signal itself was reliable: across 20 repetitions of the five-fold procedure, domain assignments matched the modal classification for an average of 96.1% of items.

C3 showed a different pattern. Overall model fit remained unchanged ($\Delta \text{CFI} = +.001$). Although then mean intercorrelations decreased slightly ($\Delta \text{Mean IC} = -.023$), Figure 1 shows increased overlap in key domain pairs, most notably between SE and CO ($.70 \rightarrow .77$; $\Delta \approx +.07$) and between SE and ER ($.62 \rightarrow .67$; $\Delta \approx +.05$). These increases suggest reduced separation between closely related domains despite stronger within-domain similarity. Figure 1 also shows that C1 reduced cross-domain correlations while preserving factor loadings, whereas C3 concentrated items in semantically shared regions, amplifying overlap among domains. The selection criteria help explain these different outcomes: C1 favors items that are closer to their intended domain than to competing domains, whereas C3 favors proximity to the target domain without directly considering proximity to

competing domains. Thus, within the diffuse SEC embedding space, maximizing own-domain similarity may preferentially select items from regions shared by closely related domains and consequently increase their empirical overlap.

Sample-level results (Tables 4 and 5) closely mirrored the median results reported in Tables 2 and 3. The only notable exception was C3 in Internet D, where the model fit declined modestly ($\Delta \text{CFI} = -.021$) rather than remaining near zero as observed in the other three samples.

	High School	College	Internet C	Internet D
$\Delta \text{CFI, C1}$	.013	.011	.040	.024
$\Delta \text{CFI, C2}$	-.004	-.029	-.002	-.033
$\Delta \text{CFI, C3}$	.005	.003	.005	-.021
$\Delta \text{Mean IC, C1}$	-.076	-.045	-.084	-.102
$\Delta \text{Mean IC, C2}$	-.053	-.020	-.040	-.067
$\Delta \text{Mean IC, C3}$	-.031	-.019	-.014	-.029

Table 4: Changes in CFI and Mean Latent Factor Intercorrelations by Sample

	High School	College	Internet C	Internet D
$\Delta \text{Load, C1}$	.008	.015	.019	.026
$\Delta \text{Load, C2}$	-.022	-.012	-.012	-.008
$\Delta \text{Load, C3}$	.005	.007	.002	.007
$\Delta \text{AVE, C1}$	.000	.011	.020	.020
$\Delta \text{AVE, C2}$	-.036	-.022	-.018	-.021
$\Delta \text{AVE, C3}$	-.002	.006	.005	.003
$\Delta \omega, \text{C1}$	.005	.009	.011	.018
$\Delta \omega, \text{C2}$	-.017	-.009	-.009	-.010
$\Delta \omega, \text{C3}$	.002	.005	.001	.004
$\Delta \alpha, \text{C1}$	-.001	.004	.007	.016
$\Delta \alpha, \text{C2}$	-.020	-.012	-.013	-.014
$\Delta \alpha, \text{C3}$	-.003	.001	-.001	.002

Table 5: Changes in Convergent Validity and Reliability by Sample

3.3 Predictive Validity

Table 6 presents adjusted R^2 from regression models predicting winter GPA, spring GPA, and academic engagement in the high school sample. Effect sizes are reported as Cohen's f^2 , as formal significance tests for differences in R^2 across non-nested models are not straightforward.

Outcome	Benchmark	C1	C2	C3
Winter GPA	.084	.092	.110	.074
Spring GPA	.030	.040	.056	.039
Academic Engagement	.324	.357	.360	.334

Table 6: Adjusted R^2 in High School Sample

Both C1 and C2 improved predictive validity relative to the benchmark across all three outcomes. For GPA outcomes, C2 produced the largest gains (winter GPA: $\Delta = +.026$, $f^2 = .029$; spring GPA: $\Delta = +.026$, $f^2 = .028$). For academic engagement, C1 ($f^2 = .051$) and C2 ($f^2 = .056$) yielded comparable improvements. All effects for C1 and C2 were in the small range. In contrast, C3 showed negligible effects. Academic engagement increased only modestly ($f^2 = .016$). Spring GPA

showed a trivial gain ($f^2 = .009$), and winter GPA declined ($f^2 = -.011$). This pattern is consistent with the elevated SE-CO latent intercorrelation observed under C3. Greater overlap among domain scores may reduce their unique contributions in joint regression models, limiting gains in predictive validity.

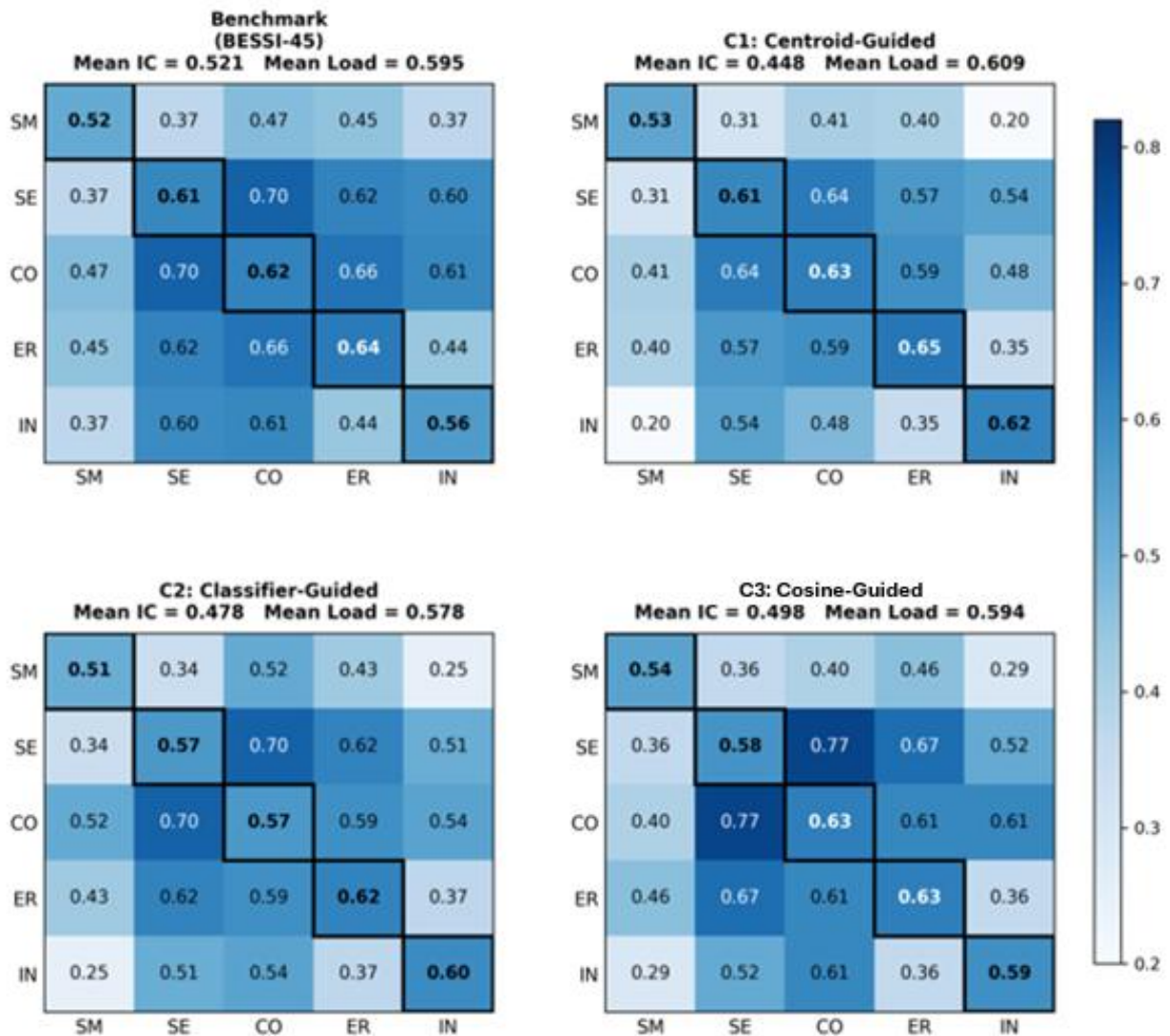


Figure 1. Latent Factor Intercorrelations (off-Diagonal) and Standardized Factor Loadings (Diagonal) by Condition

4 Discussion

Centroid margin produced the strongest overall improvement in construct validity across samples, whereas classifier confidence produced larger gains in predictive validity, particularly for GPA. These contrasting patterns suggest that geometric distinctiveness and probabilistic separability may therefore capture different aspects of item selection. Own-domain cosine similarity did not improve overall construct validity and increased overlap between some domains. When domain centroids are close, selecting items solely for similarity to their intended domain may favor content shared with competing domains. Centroid margin addresses this problem by evaluating similarity to the intended domain relative to competing domains. More broadly, the findings suggest that considering competing domains may be important for embedding-guided item selection, as both centroid margin and classifier confidence incorporate information about alternative domain memberships whereas cosine similarity does not.

These findings extend the use of embeddings from descriptive analysis to item-selection decision in instrument development. Embedding-derived diagnostics may help developers write semantically distinct items, screen candidate pools before pilot testing, and identify items likely to cross-load during revision. They may be especially useful early in development, when empirical data are unavailable and repeated pilot testing is costly.

The relatively small changes in convergent validity likely reflect the strength of the benchmark. BESSI-45 already has high reliability and adequate factor loadings, leaving limited room for further improvement in within-domain coherence through item substitution alone. Under these conditions, embedding-guided selection primarily improved discriminant validity by clarifying domain boundaries. Larger gains in convergent validity may occur with less optimized item pools.

Two limitations should be considered. First, three of the four evaluation samples – High School, College, and Internet C – also contributed response data to development of BESSI-45 (Sewell et al., 2025). This overlap may favor the benchmark and therefore understate the relative performance of the

embedding-guided conditions. C1's discriminant-validity gain over the benchmark was as large or larger in Internet D, the independent sample. Internet D showed the largest reduction in mean latent factor intercorrelation (Δ Mean IC = $-.102$). C1's improvement therefore does not appear to depend on evaluation in samples used to develop the benchmark, suggesting that the finding generalizes beyond those samples. Second, all comparisons used a single embedding model, *all-roberta-large-v1*. Future studies should examine whether the findings generalize across embedding architectures and instruments.

References

- Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., & Bong, C. H. (2014). Predicting survey responses: How and why semantics shape survey statistics on organizational behaviour. *PLoS ONE*, 9(9), e106361. <https://doi.org/10.1371/journal.pone.0106361>
- Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development*, 82(1), 405–432. <https://doi.org/10.1111/j.1467-8624.2010.01564.x>
- Chen, F. F. (2007). *Sensitivity of goodness of fit indexes to lack of measurement invariance. Structural Equation Modeling*, 14(3), 464–504.
- Cheung, G. W., & Rensvold, R. B. (2002). *Evaluating goodness-of-fit indexes for testing measurement invariance. Structural Equation Modeling*, 9(2), 233–255.
- Debelak, R., Koch, T. K., Aßenmacher, M., & Stachl, C. (2025). From embeddings to explainability: A tutorial on large-language-model-based text analysis for behavioral scientists. *Advances in Methods and Practices in Psychological Science*, 8(3), 1–24. <https://doi.org/10.1177/25152459251351285>
- Feraco, T., & Toffalini, E. (2025). SEMbeddings: How to evaluate model misfit before data collection using large-language models. *Frontiers in Psychology*, 15, 1433339. <https://doi.org/10.3389/fpsyg.2024.1433339>
- Galton, F. (1884). Measurement of character. *Fortnightly Review*, 36, 179–185.
- Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, 48(1),

- 26–34. <https://doi.org/10.1037/0003-066X.48.1.26>
- Kyllonen, P. C., & Zu, J. (2025). Assessment of social-emotional, character, behavioral, and intrapersonal and interpersonal skills. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (5th ed., pp. 1339–1421). Oxford University Press.
- 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. arXiv. <https://arxiv.org/abs/1907.11692>
- Milano, N., Luongo, M., Ponticorvo, M., & Marocco, D. (2025). Semantic analysis of test items through large language model embeddings predicts a-priori factorial structure of personality tests. *Current Research in Behavioral Sciences*, 8, 100168. <https://doi.org/10.1016/j.crbeha.2025.100168>
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1410>
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Sewell, M. N., Yoon, H. J., Lechner, C. M., Napolitano, C. M., Rammstedt, B., Roberts, B. W., & Soto, C. J. (2025). Assessing social, emotional, and behavioral skills in just a few minutes: 96-, 45-, and 20-item short forms of the BESSI. *Assessment*, 32(4), 501–520. <https://doi.org/10.1177/10731911241256434>
- Shapiro, V. B., Kim, B. K. E., Robitaille, J. L., & LeBuffe, P. A. (2024). Universal SEL screening in K–12 settings: A psychometric analysis of the DESSA-mini across three elementary school cohorts. *Assessment for Effective Intervention*, 49(2), 103–115.
- Soto, C. J., Napolitano, C. M., Sewell, M. N., Yoon, H. J., & Roberts, B. W. (2022). An integrative framework for conceptualizing and assessing social, emotional, and behavioral skills: The BESSI. *Journal of Personality and Social Psychology*, 123(1), 192–222. <https://doi.org/10.1037/pspp0000401>
- Voss, N. M., Wu, F. Y., Javalagi, A. A., & Kell, H. J. (2026). Integrating ensemble clustering and text embeddings for estimating the factor loadings of self-report scales. *Educational and Psychological Measurement*. Advance online publication. <https://doi.org/10.1177/00131644261430762>
- Wulff, D. U., & Mata, R. (2025). Escaping the jingle-jangle jungle: Increasing conceptual clarity in psychology using large language models. *Current Directions in Psychological Science*. Advance online publication. <https://doi.org/10.1177/09637214251382083>