

A Preliminary Semantic Screen for IRT Local Dependence

Qian Shen

College of Education
University of Florida
Gainesville, FL 32611, USA
qian.shen1@uf1.edu

Abstract

This work-in-progress study examines whether sentence embeddings can flag item pairs at risk for local dependence in multidimensional IRT. Using a 50-item personality inventory, preliminary analyses link semantic similarity to residual item-pair dependence. Future work will test robustness across larger samples, additional datasets, and alternative diagnostics.

1 Introduction

As MBTI, Big Five, and other personality-type testing standards have received increasingly widespread attention and recognition, people have begun to try to understand and pay sustained attention to their own corresponding personality characteristics, and they often assess their personality type by completing a wide variety of psychological test items. As with all testing items, researchers also frequently use IRT to analyze personality-type measurement items. Personality-type measurement items typically need to cover multiple dimensions; for example, MBTI test items need to cover 4 dimensions, whereas Big Five personality test items need to cover 5 dimensions. Traditional IRT assumes conditional independence among items, but in practice, this assumption is often violated due to semantic overlap among items, resulting in LD. Within each dimension and across dimensions of personality-type measurement item sets, LD may arise because of similarities in the wording and text of item stems, so that when people answer different items that appear similar, their responses are influenced by the semantic similarity among these items, and this interference can in turn affect the credibility of the resulting personality test outcomes.

2 Background

Researchers have explored more effective methods for LD detection in IRT models from a statistical-

methods perspective using multiple approaches. [Tang et al. \(2020\)](#) proposed a threshold autoregressive item response theory (TAR-IRT) model that additionally accounts for sequential dependence among each examinee's item responses. [Liu and Maydeu-Olivares \(2013\)](#) introduced a new general statistic R^2 to evaluate sources of model misfit and compared its performance with other statistics via simulation studies. [Edwards et al. \(2018\)](#) proposed a new LD diagnostic method based on parameter stability under an item-level jackknife resampling procedure.

[Wilson \(1988\)](#) proposed a method for detecting and interpreting deviations from the inter-item LD assumption for items that may share common stimulus materials or other substantive features. [Rajlic \(2019\)](#) found a significant correlation between violations of the local independence assumption defined by LD strength and the magnitude of bias in item parameter estimates. Moreover, some researchers have explored solutions when LD occurs. [Ip \(2010\)](#) argued that multidimensional response data don't necessarily require multidimensional models and discussed situations in which a unidimensional model with LD may be appropriate as an alternative. [Olsbjerg and Christensen \(2015\)](#) proposed a method for modeling LD in longitudinal 2PL models. This method is based on the concept of item splitting and enables more accurate estimation of changes in the latent variable. However, researchers have thus far primarily devoted their efforts to theoretical exploration of statistical methods, without sufficiently focusing on how to integrate the potential textual similarity among items with LD.

However, researchers have thus far primarily devoted their efforts to theoretical exploration of statistical methods, without sufficiently focusing on how to integrate the potential textual similarity among items with LD.

3 Method

3.1 Data Sources

The data I used comes from a five-point Likert Big Five personality scale that comprises 50 items and is intended to measure the five dimensions of the Big Five personality traits. Each item employed a five-point scoring format ranging from “strongly disagree” (1) to “strongly agree” (5). After missing values had been removed, 5,000 samples were selected via random sampling without replacement, and these constitute the data used in this study. This dataset serves as an initial methodological test case rather than a basis for universal screening cutoffs.

3.2 Semantic Feature Extraction

The objective of semantic feature extraction is to convert the text of the personality test items in my data into computable numerical representations, so that the semantic similarity between items can be objectively quantified. The pretrained language model Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) can map sentences of arbitrary length into a fixed-dimensional dense vector space, such that semantically similar sentences are closer together in that space. This study used the pretrained paraphrase-multilingual-MiniLM-L12-v2 model as the SBERT embedding model, extracted vector representations for the 50 items, and computed the cosine similarity between every pair of items.

3.3 Multidimensional IRT Modeling

IRT provides a rigorous measurement framework that uses latent trait models to describe the mathematical relationship between the probability of response of a respondent and their latent ability. One of the core assumptions of IRT is local independence, namely that, conditional on the respondent’s latent trait vector θ , item responses should be independent, as stated in Equation (1).

The violation of this assumption is referred to as LD and typically stems from content overlap among items, common method effects, or correlations in measurement error. Before LD detection is performed, it is essential to first establish a well-fitting IRT model because LD identification is based on residual correlations after controlling for the latent trait. If the baseline model itself fits poorly, the residuals will be confounded by model misspecification, preventing accurate identification

of true LD. In this study, I first determined the latent dimensional structure of the data using exploratory factor analysis (EFA), and the results of parallel analysis suggested extracting eight factors.

Based on the eight-dimensional structure, I fitted an eight-dimensional Graded Response Model (GRM) (Samejima, 2016). For item i , let $\mathbf{a}_i$ denote its eight-dimensional discrimination vector and b_{ik} the threshold for category boundary k . For respondent j , θ_j denotes the eight-dimensional latent trait vector. The category probability is defined in Equation (2), with $P_{ij,1}^* = 1$ and $P_{ij,K+1}^* = 0$ defining the category boundaries.

To avoid the computational challenges faced by parameter estimation in multidimensional IRT models, this study adopted the Quasi-Monte Carlo Expectation-Maximization (QMCEM) (Jank, 2005) algorithm, which uses quasi-Monte Carlo methods to approximate integration, reducing computational complexity from exponential to linear. QMCEM substantially increases computational efficiency while maintaining estimation accuracy, supporting estimation of the eight-dimensional model used in this study. The eight-dimensional GRM fitted in this study achieved excellent fit indices (CFI = 0.963, TLI = 0.948, RMSEA = 0.043), markedly superior to the unidimensional model (CFI = 0.380, TLI = 0.354, RMSEA = 0.154); this not only supports the reasonableness of the multidimensional structure but also provides a reliable basis for subsequent LD analysis.

After the model converged, I used oblimin oblique rotation to obtain a simple-structure factor loading matrix. By examining each item’s loadings on 8 dimensions, I assigned each item to the dimension with its highest loading; this operational assignment was used only to classify item pairs as within-dimension or between-dimension and does not imply exclusive loadings. The 8 dimensions each contained 2 to 8 items, and the average loading of items on their primary dimension was 0.62 (SD = 0.18). In addition, I used EAP to estimate each respondent’s latent trait scores on the 8 dimensions.

3.4 Local Dependence Detection via Partial Correlation

Traditional methods for detecting local dependence within the IRT framework include the LD χ^2 statistic proposed by (Chen and Thissen, 1997) and the (Yen, 1984) Q_3 statistic. LD χ^2 is based on deviations between observed frequencies and model-

expected frequencies, conducting a chi-square test on a two-way joint frequency table for item pairs. However, in multidimensional models, computing expected frequencies requires integration over the high-dimensional latent trait space, and as the number of categories increases, the computation becomes extremely complex. In view of the limitations of traditional methods, this study employed partial correlation to detect local dependence. Partial correlation measures the remaining linear association between two variables after controlling for a set of variables, and its theoretical basis is fully consistent with the conditional independence assumption in IRT.

Expected a posteriori (EAP) scores were estimated on the eight fitted dimensions. For each item i , responses X_i were regressed on these eight EAP scores, producing residuals e_i . For an item pair (i, j) , residual dependence was summarized by the residual correlation in Equation (3). This quantity represents the association after controlling for estimated factor scores, rather than conditioning on error-free latent traits.

To place the residual correlations on a common large-sample screening scale, I applied Fisher's transformation and defined the exploratory statistic in Equation (4). This statistic is used as a Fisher- z screening quantity rather than as the standard Yen Q_3 or Chen–Thissen LD statistic. Preliminary meaningful LD required both $|r_{ij.\theta}| > 0.10$ and $T_{ij} > 10$. In the present sample, every pair satisfying $|r_{ij.\theta}| > 0.10$ also satisfied $T_{ij} > 10$, so the effect-size criterion determined the final meaningful-LD count. Each item pair therefore had a continuous residual correlation, an exploratory T_{ij} score, and a binary preliminary meaningful-LD indicator.

3.5 Linking Semantic Similarity to Local Dependence

After completing semantic feature extraction and local dependence detection, I used correlation analysis to establish the basic association between semantic similarity and LD. I computed Pearson and Spearman correlations. To describe heterogeneity by dimension status, I computed Pearson and Spearman correlations separately for within-dimension and between-dimension item pairs. Because the same item appears in multiple pairs, these pair-level associations were treated as descriptive rather than as independence-adjusted inferential tests.

Second, hierarchical regression analysis was

used to describe incremental associations and interaction patterns. The dependent variable was the absolute residual correlation $|r_{ij.\theta}|$. I specified three nested models: Model 1 included only semantic similarity; Model 2 added same-dimension status; and Model 3 added their interaction. The three models are given in Equations (5)–(7). I conducted a simple slope analysis to describe the fitted effect of semantic similarity under the two dimension-status conditions: the within-dimension slope is $\beta_1 + \beta_3$, whereas the between-dimension slope is β_1 . Changes in R^2 across the nested models were treated descriptively because item pairs share items and therefore do not constitute independent observations.

Finally, a secondary logistic regression used the preliminary meaningful-LD indicator (0/1) as the dependent variable. The model included semantic similarity, same-dimension status, and their interaction, as shown in Equation (8). Model fit evaluation used $-2 \log L$, Nagelkerke pseudo- R^2 (Walker and Smith, 2016), and the Hosmer–Lemeshow goodness-of-fit test (Paul et al., 2013).

4 Results

4.1 Local Dependence Patterns

After controlling for EAP estimates of the eight fitted dimensions, the residual correlations across the 1,225 item pairs had a median of -0.002 (IQR = $[-0.023, 0.021]$), indicating that most pairwise residual associations were close to zero. Using the preliminary meaningful-LD criterion, 120 of 1,225 pairs (9.8%) were flagged. As shown in Table 1, 52.8% of within-dimension pairs and 2.4% of between-dimension pairs met this criterion. The mean absolute residual correlation was also larger for within-dimension pairs ($M = 0.091$, $SD = 0.087$) than for between-dimension pairs ($M = 0.023$, $SD = 0.037$). These pair-level differences are reported descriptively because item pairs are not independent.

4.2 Semantic Similarity Predicts Local Dependence

Semantic similarity showed a preliminary positive association with absolute residual dependence. Across all item pairs, the Pearson correlation between semantic similarity and $|r_{ij.\theta}|$ was $r = 0.500$, while the Spearman correlation was $\rho = 0.271$. Semantic similarity alone accounted for 25.0% of the pair-level variance ($R^2 = 0.250$).

As shown in Table 2, the association was stronger within dimensions ($r = 0.651$, $R^2 = 42.4\%$) than between dimensions ($r = 0.209$, $R^2 = 4.4\%$). Thus, semantic similarity was informative within the operational dimension groups but comparatively weak across groups. These pair-level estimates are descriptive because individual items occur in multiple pairs.

4.3 Hierarchical Regression

Table 3 summarizes the descriptive hierarchical regressions. Semantic similarity alone accounted for $R^2 = 0.250$ of the pair-level variance. Adding same-dimension status increased R^2 to 0.486, and adding the interaction increased it to 0.585 ($\Delta R^2 = 0.100$). As shown in Table 4, the fitted interaction coefficient was $B = 0.386$. The corresponding fitted semantic-similarity slope was 0.054 for between-dimension pairs and 0.440 for within-dimension pairs. These estimates indicate a stronger fitted association within dimensions, but the pair-level uncertainty estimates are not adjusted for repeated item membership.

4.4 Logistic Regression

Table 5 summarizes a secondary logistic regression using the preliminary meaningful-LD indicator as the outcome. Semantic similarity and same-dimension status had positive fitted coefficients, whereas the interaction coefficient was not distinguishable from zero in the unadjusted pair-level analysis ($p = 0.401$). The Nagelkerke pseudo- R^2 was 0.456. Because the same item appears in multiple pairs and the meaningful-LD prevalence was 9.8%, this model is treated as descriptive and in-sample; practical screening performance requires out-of-sample validation.

5 Conclusion

This preliminary analysis suggests that semantic similarity contains useful information about residual item dependence, especially for items assigned to the same dimension. It does not establish a general screening rule and does not replace psychometric calibration or content review. At this stage, the method is best viewed as a way to prioritize item pairs for further investigation.

Because this is an initial analysis of a 5,000-responder sample from one dataset, robustness across datasets and sample sizes has not yet been established. Future work will evaluate additional

datasets, sample sizes, text representations, established LD diagnostics, and dependence-adjusted inference.

References

- Wen-Hung Chen and David Thissen. 1997. Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22(3):265–289.
- Michael C Edwards, Carrie R Houts, and Li Cai. 2018. A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological methods*, 23(1):138.
- Edward Haksing Ip. 2010. Empirically indistinguishable multidimensional irt and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology*, 63(2):395–416.
- Wolfgang Jank. 2005. Quasi-monte carlo sampling to improve the efficiency of monte carlo em. *Computational statistics & data analysis*, 48(4):685–701.
- Yang Liu and Alberto Maydeu-Olivares. 2013. Local dependence diagnostics in irt modeling of binary data. *Educational and Psychological Measurement*, 73(2):254–274.
- Maja Olsbjerg and Karl Bang Christensen. 2015. Modeling local dependence in longitudinal irt models. *Behavior research methods*, 47(4):1413–1424.
- Prabasaj Paul, Michael L Pennell, and Stanley Lemeshow. 2013. Standardizing the power of the hosmer–lemeshow goodness of fit test in large data sets. *Statistics in medicine*, 32(1):67–80.
- Gordana Rajlic. 2019. *Violations of unidimensionality and local independence in measures intended as unidimensional: assessing levels of violations and the accuracy in unidimensional IRT model estimates*. Ph.D. thesis, University of British Columbia.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
- Fumiko Samejima. 2016. Graded response models. In *Handbook of item response theory*, pages 95–107. Chapman and Hall/CRC.
- Xiaodan Tang, George Karabatsos, and Haiqin Chen. 2020. Detecting local dependence: A threshold-autoregressive item response theory (tar-irt) approach for polytomous items. *Applied Measurement in Education*, 33(4):280–292.
- David A Walker and Thomas J Smith. 2016. Nine pseudo r2 indices for binary logistic regression models. *Journal of Modern Applied Statistical Methods*, 15(1):848–854.

- Mark Wilson. 1988. Detecting and interpreting local item dependence using a family of Rasch models. *Applied psychological measurement*, 12(4):353–364.
- Wendy M Yen. 1984. Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8(2):125–145.

A Equations

$$P(\mathbf{X} \mid \boldsymbol{\theta}) = \prod_{i=1}^I P(X_i \mid \boldsymbol{\theta}). \quad (1)$$

$$P(X_{ij} = k \mid \boldsymbol{\theta}_j) = P_{ij,k}^* - P_{ij,k+1}^*, \quad P_{ij,k}^* = \text{logit}^{-1}(\mathbf{a}_i^\top \boldsymbol{\theta}_j - b_{ik}). \quad (2)$$

$$r_{ij \cdot \boldsymbol{\theta}} = \text{cor}(e_i, e_j). \quad (3)$$

$$z_{ij}^{(F)} = \text{atanh}(r_{ij \cdot \boldsymbol{\theta}}), \quad T_{ij} = N \left[z_{ij}^{(F)} \right]^2. \quad (4)$$

$$|r_{ij \cdot \boldsymbol{\theta}}| = \beta_0 + \beta_1 \cdot \text{semantic_sim}_{ij} + \epsilon_{ij}. \quad (5)$$

$$|r_{ij \cdot \boldsymbol{\theta}}| = \beta_0 + \beta_1 \cdot \text{semantic_sim}_{ij} + \beta_2 \cdot \text{same_dimension}_{ij} + \epsilon_{ij} \quad (6)$$

$$|r_{ij \cdot \boldsymbol{\theta}}| = \beta_0 + \beta_1 \cdot \text{semantic_sim}_{ij} + \beta_2 \cdot \text{same_dimension}_{ij} + \beta_3 \cdot (\text{semantic_sim}_{ij} \times \text{same_dimension}_{ij}) + \epsilon_{ij} \quad (7)$$

$$\begin{aligned} \log \left(\frac{P(\text{LD}_{ij} = 1)}{1 - P(\text{LD}_{ij} = 1)} \right) &= \gamma_0 + \gamma_1 \cdot \text{semantic_sim}_{ij} + \gamma_2 \cdot \text{same_dimension}_{ij} \\ &+ \gamma_3 \cdot (\text{semantic_sim}_{ij} \times \text{same_dimension}_{ij}) \end{aligned} \quad (8)$$

B Tables

Table 1: Local Dependence by Dimensional Structure

Criterion	Within-Dimension ($n = 180$)	Between-Dimension ($n = 1,045$)	Ratio
Fisher- z screen ($T_{ij} > 10$)	148 (82.2%)	171 (16.4%)	5.0×
Residual effect size ($ r_{ij,\theta} > 0.10$)	95 (52.8%)	25 (2.4%)	22.0×
Preliminary meaningful LD (both criteria)	95 (52.78%)	25 (2.39%)	22.1×

Table 2: Correlation between Semantic Similarity and Local Dependence by Item Pair Type

Item Pair Type	n	Semantic vs $ r_{\text{partial}} $	Semantic vs T_{ij} r	Variance Explained
Within-dimension	180	$r = 0.651$	$r = 0.614$	42.4%
Between-dimension	1,045	$r = 0.209$	$r = 0.225$	4.4%

Note. Values are descriptive pair-level estimates; no independence-adjusted significance tests are reported.

Table 3: Hierarchical Regression Predicting Local Dependence Intensity

Model	Predictors	R^2	ΔR^2
1	Semantic similarity	0.250	—
2	+ Same dimension	0.486	0.236
3	+ Semantic $\times$ Dimension	0.585	0.100

Note. $n = 1,225$ item pairs. Estimates are descriptive because item pairs share items.

Table 4: Regression Coefficients for Model 3 (Interaction Model)

Predictor	B	SE	β	t	p
Intercept	0.013	0.003	—	4.48	< 0.001
Semantic similarity	0.054	0.011	0.104	5.95	< 0.001
Same dimension	−0.047	0.008	−0.141	−5.73	< 0.001
Semantic $\times$ Dimension	0.386	0.022	0.548	17.14	< 0.001

Note. SE, t , and p values are unadjusted for repeated item membership and are reported as exploratory pair-level statistics.

Table 5: Logistic Regression Predicting Meaningful Local Dependence

Predictor	B	SE	z	p
(Intercept)	−6.448	0.663	−9.719	< 0.001
Semantic similarity	9.408	1.850	5.084	< 0.001
Same dimension	3.935	0.861	4.572	< 0.001
Semantic $\times$ Dimension	−2.011	2.396	−0.839	0.401

Note. SE, z , and p values are unadjusted for repeated item membership and are reported as exploratory pair-level statistics.

C Data Availability

Link of the data <https://www.kaggle.com/datasets/tunguz/big-five-personality-test>