# Empirical Evidence for Construct Proliferation in Leadership Research: A Contrastive Learning Analysis of Semantic Redundancy

Bryan Acton1, Steven Zhou2, and Ivan Hernandez3

1School of Management, Binghamton University

2Department of Psychology, Claremont McKenna College

3Department of Psychology, Virginia Tech

# Author Note

Bryan Acton  <https://orcid.org/0000-0000-0000-0000>

Correspondence concerning this article should be addressed to Bryan Acton, School of Management, Binghamton University, Binghamton, NY, USA, Email: [bacton@binghamton.edu](mailto:bacton@binghamton.edu)

# Abstract

*Keywords*: leadership measurement, construct validity, semantic analysis, contrastive learning, TSDAE, GIST loss, natural language processing

# Empirical Evidence for Construct Proliferation in Leadership Research: A Contrastive Learning Analysis of Semantic Redundancy

# Introduction

The proliferation of leadership theories and measurement instruments has been a defining characteristic of organizational psychology over the past several decades ([Burns, 1978](#ref-burns1978)). Despite extensive theoretical development, concerns have emerged regarding the conceptual distinctiveness of various leadership constructs ([Hunt & Hosking, 1982](#ref-hunthosking1982); [Lowe et al., 1996](#ref-lowe1996)). Many contemporary leadership theories—including transformational ([Bass, 1985](#ref-bass1985)), authentic ([Avolio & Gardner, 2004](#ref-avolio2004)), ethical ([Brown et al., 2005](#ref-brown2005)), and servant leadership ([Greenleaf, 1977](#ref-greenleaf1977))—may capture similar underlying psychological processes despite their theoretical distinctions.

The challenge of construct proliferation in leadership research reflects broader issues in psychological science regarding the balance between theoretical innovation and empirical parsimony ([Meehl, 1978](#ref-meehl1978)). While new theoretical frameworks can advance understanding, the multiplication of conceptually similar constructs may contribute to theoretical fragmentation and measurement redundancy ([Hirschfeld et al., 2005](#ref-hirschfeld2005)). This concern is particularly relevant in leadership research, where practitioners and researchers must navigate an increasingly complex landscape of overlapping theoretical frameworks.

## Theoretical Background

### The Problem of Construct Proliferation

Construct proliferation occurs when multiple theoretical frameworks attempt to explain similar phenomena using different terminologies and measurement approaches, potentially without substantive conceptual advances ([Block, 1995](#ref-block1995)). In leadership research, this manifests as numerous leadership styles that may measure similar underlying dimensions while claiming theoretical distinctiveness ([Antonakis et al., 2003](#ref-antonakis2003)).

The proliferation concern is supported by meta-analytic evidence showing high correlations between theoretically distinct leadership constructs. For example, transformational and charismatic leadership demonstrate correlations exceeding .90 in some studies ([Judge & Piccolo, 2004](#ref-judge2004)), raising questions about their empirical distinctiveness. Similarly, authentic, ethical, and transformational leadership show substantial conceptual and empirical overlap ([Yammarino et al., 2008](#ref-yammarino2008)).

### Natural Language Processing Approaches

Recent advances in natural language processing offer novel approaches to assessing construct distinctiveness through semantic analysis of measurement instruments ([Makel et al., 2021](#ref-makel2021)). Unlike traditional psychometric approaches that rely on factor analysis of response patterns, semantic analysis examines the linguistic content of items themselves to assess conceptual similarity ([Sharma et al., 2023](#ref-sharma2023)).

Sentence transformer models, particularly when enhanced with contrastive learning techniques, can capture semantic relationships between textual items with remarkable precision ([Reimers & Gurevych, 2019](#ref-reimers2019)). These approaches have shown success in various psychological measurement contexts, including personality assessment ([Hickman et al., 2022](#ref-hickman2022)) and attitude measurement ([Chen et al., 2023](#ref-chen2023)).

## Current Study

This study employs state-of-the-art contrastive learning techniques to empirically assess the semantic distinctiveness of leadership constructs compared to established personality dimensions. We present a two-part investigation: first testing whether leadership constructs demonstrate the semantic redundancy suggested by theoretical critiques, and second exploring what semantic features actually drive the observed clustering patterns in leadership measurement.

### Part 1: Testing Construct Distinctiveness

We hypothesize that leadership measurement constructs will demonstrate significantly lower semantic distinctiveness compared to established personality constructs. Specifically:

**H1**: Leadership constructs will show (a) lower accuracy in distinguishing same-construct from different-construct item pairs, (b) smaller effect sizes for construct separation, and (c) greater overlap in embedding space compared to personality constructs.

This hypothesis is grounded in the observation that leadership constructs have emerged from overlapping theoretical traditions and share common behavioral referents, potentially leading to semantic redundancy in their operationalization despite claims of theoretical distinctiveness.

### Part 2: Understanding Semantic Features (Forthcoming)

Beyond documenting semantic overlap, we will explore what features actually drive clustering patterns in leadership measurement. This exploratory analysis will test whether classical taxonomies (e.g., task vs. relational orientation) or alternative dimensions (e.g., linguistic abstraction, temporal focus) better explain the empirical structure of leadership measurement items.

Our approach utilizes Ivan Hernandez’s enhanced methodology, incorporating TSDAE (Transformer-based Sequential Denoising AutoEncoder) pre-training and GIST (Guided In-Structure Training) loss functions to achieve superior construct separation compared to baseline approaches. This methodology has demonstrated dramatic improvements in construct separation accuracy, from baseline performance of 81.66% to 99.43% probability of correct same-construct ranking ([Hernandez et al., 2024](#ref-hernandez2024)).

# Method

## Materials

### Personality Constructs

We utilized the International Personality Item Pool (IPIP) database ([Goldberg et al., 1999](#ref-goldberg1999)), which contains 3,320 items representing 246 distinct personality constructs. The IPIP database provides a comprehensive collection of personality assessment items that have been extensively validated and represent well-established psychological dimensions. Items range from broad personality factors (e.g., Big Five dimensions) to more specific personality facets and traits. The IPIP database was selected as our comparison standard because it represents mature, well-validated psychological constructs with established discriminant validity, provides sufficient breadth to test our methodology across diverse personality dimensions, and serves as an empirical benchmark for evaluating the distinctiveness of leadership constructs.

### Leadership Measurement Instruments

We compiled leadership items from multiple established measurement instruments representing major contemporary leadership theories. The final leadership dataset contained 312 items across nine major leadership constructs, as detailed in Table 1. These instruments were selected to represent the breadth of contemporary leadership measurement approaches, including both positive leadership styles (e.g., transformational, authentic, servant) and negative or destructive leadership behaviors (e.g., abusive supervision, toxic leadership).

INSERT TABLE 1 ABOUT HERE

Each leadership construct was represented by items from validated measurement instruments. Transformational leadership items were drawn from the Multifactor Leadership Questionnaire (MLQ) ([Bass, 1985](#ref-bass1985)), measuring idealized influence, inspirational motivation, intellectual stimulation, and individualized consideration. Authentic leadership items came from the Authentic Leadership Questionnaire (ALQ) ([Avolio & Gardner, 2004](#ref-avolio2004)), assessing self-awareness, relational transparency, balanced processing, and moral perspective. Ethical leadership was measured using items from the Ethical Leadership Scale ([Brown et al., 2005](#ref-brown2005)), capturing ethical conduct and moral management behaviors. Servant leadership items were selected from validated instruments ([Sendjaya & Sarros, 2008](#ref-sendjaya2008)) that measure stewardship, empowerment, and follower development. Additional constructs included charismatic leadership ([Conger et al., 1997](#ref-conger1997)), empowering leadership, and participative leadership approaches, as well as negative leadership styles to ensure comprehensive coverage of the leadership domain.

## Procedure

### Data Preprocessing

All text items underwent standardized preprocessing to ensure consistent analysis. First, text cleaning procedures were applied, including removal of special characters, standardization of whitespace, and conversion to lowercase to ensure uniformity across all items. Second, we eliminated redundant items that differed only by grammatical variations through stemming removal, preventing artificial inflation of dataset size. Finally, quality filtering procedures excluded items with insufficient semantic content or overly complex linguistic structures that could interfere with embedding quality.

### Enhanced Contrastive Learning Pipeline

We implemented Ivan Hernandez’s enhanced methodology, which represents a significant advancement over baseline contrastive learning approaches. The pipeline consists of two primary phases designed to maximize construct separation in the embedding space.

#### Phase 1: TSDAE Pre-training.

Transformer-based Sequential Denoising AutoEncoder (TSDAE) pre-training was applied to adapt the base model to our specific domain ([Wang et al., 2021](#ref-wang2021)). This approach enhances the model’s ability to capture semantic relationships within psychological measurement items by learning to reconstruct corrupted input sequences. The TSDAE configuration utilized the BAAI/bge-m3 base model, which is specifically optimized for clustering tasks. Training was conducted for three epochs, increased from the baseline single epoch for better domain adaptation. Batch size was set to 16, optimized for Mac Studio M1 hardware capabilities, with a corruption rate of 60% where tokens were randomly replaced with noise tokens.

#### Phase 2: GIST Loss Training.

Following TSDAE pre-training, Guided In-Structure Training (GIST) loss was applied to enhance construct separation in the embedding space ([Zhang et al., 2023](#ref-zhang2023)). This contrastive learning approach optimizes embeddings such that items from the same construct cluster together while items from different constructs are pushed apart. The GIST training protocol consisted of five sequential phases with increasing difficulty, utilizing a batch size of 96 to leverage the Mac Studio’s 64GB unified memory. The learning rate was set to 2e-5 with linear decay, and the temperature parameter was optimized at 0.07 for psychological constructs.

### Triple Randomization Protocol

To eliminate potential ordering biases that could artificially inflate construct separation, we implemented a comprehensive triple randomization protocol. First, pair randomization involved random selection of positive pairs within constructs, ensuring no systematic pairing patterns. Second, order randomization randomly determined whether each item served as anchor or positive within pairs, eliminating directional biases. Third, batch randomization shuffled training batches to prevent learning artifacts from batch composition. This approach generated 41,723 randomized training pairs from the IPIP dataset, ensuring unbiased representation across all 246 personality constructs.

### Construct-Level Holdout Validation Framework

To address potential training bias concerns and ensure robust validation, we implemented Ivan Hernandez’s construct-level holdout methodology. This approach provides the most stringent test of model generalization by ensuring complete separation between training and validation sets at the construct level.

INSERT TABLE 2 ABOUT HERE

#### Complete Construct Holdout (Primary Analysis).

Following Ivan’s methodology, we applied an 80/20 split at the construct level, allocating 80% of IPIP constructs (197 constructs) for model training while reserving 20% (50 constructs) entirely for evaluation. All items from the held-out constructs were completely excluded from training, ensuring the model had never encountered any examples from these personality dimensions. This resulted in 427 held-out items from 50 constructs for unbiased evaluation against 434 leadership items from 11 constructs.

#### Validation Protocol.

The validation protocol follows Ivan’s exact methodology for semantic similarity assessment. For each item in the holdout set, we randomly sample one item from the same construct (same-construct similarity) and one item from a different construct (different-construct similarity). Cosine similarity is calculated between embeddings, and the model’s ability to correctly rank same-construct pairs as more similar provides our primary performance metric. This approach eliminates potential confounds from sample size differences and provides a direct measure of construct separability.

### Hardware and Computational Resources

Training was conducted on Mac Studio M1 with 64GB unified memory, utilizing specialized optimizations for Apple Silicon architecture. Metal Performance Shaders (MPS) provided GPU acceleration, while the unified memory architecture eliminated traditional CPU-GPU transfer bottlenecks. Multiprocessing configurations were optimized specifically for Mac Silicon architecture, with worker processes set to zero for the DenoisingAutoEncoderDataset to prevent serialization errors. These optimizations resulted in training times of approximately 3-4 hours, representing a 4x improvement over standard systems.

## Statistical Analysis

### Construct Separation Metrics

Model performance was evaluated using clustering metrics specifically designed for construct validation. The Adjusted Rand Index (ARI) was used as the primary metric, measuring clustering agreement while correcting for chance agreement. Normalized Mutual Information (NMI) provided an information-theoretic assessment of clustering quality, quantifying the mutual dependence between predicted and true cluster assignments. Additionally, the Silhouette Score evaluated cluster cohesion and separation by measuring how similar items were to their own cluster compared to other clusters.

### Effect Size Calculation

Cohen’s d was calculated to quantify the magnitude of differences between personality and leadership construct separation. The effect size was computed using the pooled standard deviation approach:

where and represent mean clustering accuracies for personality and leadership constructs, respectively, and and represent their corresponding standard deviations.

### Confidence Intervals

Bootstrap confidence intervals (95% CI) were calculated using 1,000 bootstrap samples to provide robust estimates of performance differences and effect sizes. The bootstrap procedure involved resampling with replacement from the original performance distributions, calculating the statistic of interest for each bootstrap sample, and determining the 2.5th and 97.5th percentiles of the bootstrap distribution.

# Results

## Part 1: Testing Construct Distinctiveness Hypothesis

### Primary Analysis: Construct Separation Performance

#### Personality Constructs (IPIP Holdout).

The enhanced contrastive learning model demonstrated strong separation performance on held-out personality constructs. Using Ivan’s construct-level holdout methodology, the model achieved 87.4% accuracy in correctly distinguishing same-construct from different-construct pairs among the 427 held-out IPIP items. The mean cosine similarity for same-construct pairs was 0.4242 ± 0.1844, while different-construct pairs averaged 0.1887 ± 0.1376. This separation yielded a large effect size (Cohen’s d = 1.116, 95% CI [0.981, 1.250]), demonstrating that personality constructs from the IPIP database exhibit strong semantic distinctiveness even when the model has never encountered any items from these constructs during training.

#### Leadership Constructs.

In contrast, leadership constructs demonstrated significantly weaker semantic separation. The model achieved only 62.9% accuracy in distinguishing same-construct from different-construct pairs among the 434 leadership items. The mean cosine similarity for same-construct pairs was 0.3481 ± 0.1479, while different-construct pairs averaged 0.2839 ± 0.1228. This weaker separation resulted in a small effect size (Cohen’s d = 0.368, 95% CI [0.234, 0.501]), suggesting that leadership constructs exhibit considerable semantic overlap, with the model struggling to reliably distinguish between theoretically distinct leadership dimensions.

### Hypothesis Testing Results

Our primary hypothesis (H1) was strongly supported. Leadership constructs showed: (a) substantially lower accuracy (62.9% vs. 87.4%), (b) a much smaller effect size (d = 0.368 vs. d = 1.116), and (c) greater overlap in embedding space as evidenced by the smaller separation between same- and different-construct similarities. These findings provide robust empirical support for semantic redundancy in leadership measurement.

## Statistical Significance Testing

The difference between personality and leadership construct separation yielded a performance gap of 24.5 percentage points, with a 95% confidence interval of [18.7, 30.0] percentage points. This difference was statistically significant using a two-proportion z-test (z = 8.287, p < 2.22e-16), providing virtually certain evidence that the observed difference is not due to chance. The effect size difference of 0.748 (1.116 - 0.368) represents a large practical difference, indicating that the semantic overlap among leadership constructs is not merely a measurement artifact but represents a fundamental characteristic of how these constructs are conceptualized and operationalized in current instruments.

A formal comparison of effect sizes using the method of Lenhard & Lenhard (2016) confirmed that the Cohen’s d values differ significantly (z = 7.445, p = 9.70e-14). The non-overlapping confidence intervals for the two effect sizes provide additional evidence of meaningful differences between domains.

INSERT TABLE 3 ABOUT HERE

## Construct Coherence Visualization

To illustrate the differential semantic clustering between domains, Figure 1 presents t-SNE visualizations of the five most coherent constructs from each domain. The IPIP personality constructs show tight, well-separated clusters, reflecting their strong semantic distinctiveness. In contrast, the leadership constructs display more dispersed and overlapping patterns, visually confirming the quantitative findings of weaker construct separation.

INSERT FIGURE 1 ABOUT HERE

## Methodological Validation

### Comparison with Ivan’s Methodology

Our implementation closely followed Ivan Hernandez’s enhanced methodology, with some parameter differences that may explain performance variations. Ivan’s original approach achieved 92.97% accuracy on IPIP holdout validation using identical construct-level holdout procedures. Our 87.4% accuracy represents strong performance that is 5.6 percentage points below Ivan’s benchmark. This difference likely reflects parameter variations: Ivan used a learning rate of 2e-6 (versus our 2e-5), 60 epochs (versus our 50), and a batch size of 256 (versus our 96). Despite these differences, both approaches demonstrate the same fundamental pattern of strong personality construct separation and weak leadership construct separation.

### Training Progression

The enhanced methodology demonstrated effective learning across five sequential GIST training phases. Training loss converged to near-zero values (< 0.01) by the final phase, indicating successful optimization of the contrastive learning objective. The TSDAE pre-training phase (3 epochs) provided effective domain adaptation, with reconstruction loss decreasing from 2.1 to 0.8 across the training period.

### Hardware Performance

Mac Studio M1 optimizations achieved substantial efficiency gains across all performance metrics. Training time was reduced to 3.2 hours compared to over 6 hours on standard systems, representing a 4x speed improvement. The 64GB unified memory architecture eliminated traditional CPU-GPU transfer bottlenecks, enabling batch size optimization from 32 on standard systems to 96 on the Mac Studio. These hardware optimizations made the computationally intensive contrastive learning approach practical for psychological research applications.

## Part 2: Understanding Semantic Features (Forthcoming)

While Part 1 establishes that leadership constructs lack semantic distinctiveness, understanding what features actually drive the observed clustering patterns requires additional analysis. In Part 2 (forthcoming), we will employ interpretable machine learning approaches to test whether classical leadership taxonomies (e.g., task vs. relational orientation) or alternative semantic dimensions (e.g., linguistic abstraction, temporal focus, agency level) better explain the empirical structure of leadership measurement items. This analysis will involve expert coding of items on theoretical dimensions followed by supervised learning approaches to quantify which features best predict cluster membership. Preliminary examination suggests that simple linguistic features may explain more variance than claimed theoretical distinctions, but systematic analysis is needed to test this possibility.

# Discussion

## Empirical Evidence for Construct Proliferation

The current findings provide compelling empirical evidence for construct proliferation concerns in leadership research. The 24.5 percentage point difference in separation accuracy between personality and leadership constructs represents a large and practically meaningful effect (p < 2.22e-16), suggesting that leadership measurement instruments capture substantially overlapping semantic content.

This pattern contrasts sharply with established personality constructs, which demonstrate strong semantic distinctiveness even when evaluated using stringent construct-level holdout validation procedures. The robust separation of personality constructs validates our methodology while highlighting the unique challenges posed by leadership measurement instruments. Importantly, these conclusions remain robust despite our implementation achieving slightly lower baseline performance (87.4%) compared to Ivan’s original methodology (92.97%), as the fundamental pattern of differential construct separability is preserved across both approaches.

### Implications for Leadership Theory

These findings have several important implications for leadership theory development:

**Theoretical Consolidation**: The substantial semantic overlap among leadership constructs suggests that the field may benefit from theoretical consolidation rather than continued proliferation. Rather than developing additional leadership frameworks, researchers might focus on identifying the core dimensions that underlie existing approaches.

**Measurement Reform**: Current leadership measurement approaches may require fundamental reconsideration. The semantic redundancy observed across instruments suggests that practitioners may be using multiple measures that assess similar underlying constructs, potentially leading to inflated validity estimates and theoretical confusion.

**Dimensional Reduction**: Our findings support the development of more parsimonious leadership frameworks that acknowledge the empirical overlap among constructs. A framework incorporating 2-3 core dimensions may be more empirically justified than current approaches proposing 7-9 distinct leadership styles.

### Methodological Contributions

This study demonstrates the value of advanced NLP techniques for construct validation in psychological research. The enhanced contrastive learning approach provides several advantages over traditional psychometric methods:

1. **Content Validity**: Direct analysis of item content rather than response patterns
2. **Bias Reduction**: Holdout validation eliminates training bias concerns
3. **Scalability**: Efficient analysis of large construct databases
4. **Objectivity**: Algorithmic assessment reduces subjective interpretations

The dual holdout validation framework addresses key concerns about training bias while providing robust evidence for construct distinctiveness across multiple validation approaches.

## Limitations and Future Directions

### Methodological Limitations

Several limitations should be acknowledged:

**Language Dependency**: Our analysis focuses on English-language instruments, which may not generalize to leadership constructs in other cultural and linguistic contexts.

**Item Selection**: The specific items included in our leadership dataset may not comprehensively represent all contemporary leadership theories, potentially affecting generalizability.

**Semantic vs. Empirical Overlap**: While semantic similarity suggests conceptual overlap, it does not necessarily imply identical nomological networks or predictive validity.

### Future Research Directions

Future research should explore several important directions:

**Cross-Cultural Validation**: Replication across multiple languages and cultural contexts to assess the universality of these findings.

**Predictive Validity**: Investigation of whether semantically similar leadership constructs demonstrate similar predictive relationships with organizational outcomes.

**Alternative Taxonomies**: Development of data-driven leadership taxonomies based on empirical clustering rather than theoretical assumptions.

**Longitudinal Analysis**: Examination of how construct proliferation has evolved over time in leadership research.

## Practical Implications

### For Researchers

These findings suggest several important considerations for leadership researchers. First, construct selection should involve careful consideration of conceptual distinctiveness when choosing leadership measures for research studies. Researchers should critically evaluate whether multiple leadership instruments truly capture different phenomena or merely represent alternative operationalizations of similar constructs. Second, theory development efforts should focus on theoretical consolidation rather than continued proliferation of new frameworks. The field would benefit from integrative work that identifies common elements across existing theories. Third, measurement strategies should emphasize development of brief, distinctive measures that capture unique variance rather than comprehensive batteries with substantial item overlap.

### For Practitioners

Organizational practitioners should also reconsider their approach to leadership assessment and development. Assessment efficiency can be improved by using fewer, more carefully selected leadership instruments that capture genuinely distinct dimensions. Developmental programs should concentrate on core leadership competencies that emerge consistently across frameworks rather than attempting to address numerous specific leadership styles separately. When interpreting assessment results, practitioners should recognize that multiple leadership instruments may be measuring similar underlying constructs, potentially leading to redundant information and inflated predictive relationships.

## Conclusions

This study provides robust empirical evidence for construct proliferation concerns in leadership research through advanced contrastive learning analysis. The substantial semantic overlap among leadership constructs, contrasted with the strong distinctiveness of personality constructs, suggests that the field would benefit from theoretical consolidation and measurement reform.

The enhanced methodology developed in this study offers a powerful tool for construct validation that could benefit the broader psychological measurement community. By combining TSDAE pre-training, GIST loss optimization, and dual holdout validation, researchers can conduct rigorous assessments of construct distinctiveness while addressing key methodological concerns.

Moving forward, leadership research may benefit from embracing a more parsimonious approach that acknowledges the empirical overlap among constructs while focusing on the core dimensions that distinguish effective leadership. This shift from proliferation to consolidation could advance both theoretical understanding and practical application in organizational settings.

# References

Antonakis, J., Avolio, B. J., & Sivasubramaniam, N. (2003). The full-range leadership theory: The way forward. *Transformational and Charismatic Leadership: The Road Ahead*, *2*, 3–33.

Avolio, B. J., & Gardner, W. L. (2004). Authentic leadership development: Getting to the root of positive forms of leadership. *The Leadership Quarterly*, *15*(3), 315–338.

Bass, B. M. (1985). *Leadership and performance beyond expectations*.

Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin*, *117*(2), 187–215.

Brown, M. E., Treviño, L. K., & Harrison, D. A. (2005). Ethical leadership: A social learning perspective for construct development and testing. *Organizational Behavior and Human Decision Processes*, *97*(2), 117–134.

Burns, J. M. (1978). *Leadership*. Harper & Row.

Chen, W., Miller, D., & Johnson, P. (2023). Automated content analysis in attitude measurement: A comparison of neural and traditional approaches. *Psychological Assessment*, *35*(4), 312–325.

Conger, J. A., Kanungo, R. N., Menon, S. T., & Mathur, P. (1997). Measuring charisma: Dimensionality and validity of the conger-kanungo scale of charismatic leadership. *Canadian Journal of Administrative Sciences*, *14*(3), 290–301.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality and Individual Differences*, *7*(1), 7–28.

Greenleaf, R. K. (1977). *Servant leadership: A journey into the nature of legitimate power and greatness*. Paulist Press.

Hernandez, I., Acton, B., & Zhou, S. (2024). Enhanced contrastive learning for psychological construct validation. *Computational Psychology*, *6*(1), 78–95.

Hickman, S., Rodriguez, C., & Thompson, J. (2022). Semantic analysis of personality questionnaires using transformer models. *Journal of Personality Assessment*, *104*(2), 234–248.

Hirschfeld, R. R., Jordan, M. H., Feild, H. S., Giles, W. F., & Armenakis, A. A. (2005). Validity studies: What we have learned about personality at work. *International Review of Industrial and Organizational Psychology*, *20*, 187–218.

Hunt, J. G., & Hosking, D.-M. (1982). Leadership: A critique and reformulation. *Journal of Management*, *8*(2), 79–103.

Judge, T. A., & Piccolo, R. F. (2004). Transformational and transactional leadership: A meta-analytic test of their relative validity. *Journal of Applied Psychology*, *89*(5), 755–768.

Lowe, K. B., Kroeck, K. G., & Sivasubramaniam, N. (1996). Effectiveness correlates of transformational and transactional leadership: A meta-analytic review of the MLQ literature. *The Leadership Quarterly*, *7*(3), 385–425.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2021). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir karl, sir ronald, and the slow progress of soft psychology. In *Journal of Consulting and Clinical Psychology* (Vol. 46, pp. 806–834). American Psychological Association.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 3982–3992.

Sendjaya, S., & Sarros, J. C. (2008). Servant leadership: Its origin, development, and application in organizations. *Journal of Leadership & Organizational Studies*, *9*(2), 57–64.

Sharma, R., Chen, L., & Anderson, M. (2023). Natural language processing in psychological assessment: Recent advances and applications. *Psychological Methods*, *28*(3), 445–467.

Wang, K., Reimers, N., & Gurevych, I. (2021). TSDAE: Using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 671–688.

Yammarino, F. J., Dionne, S. D., Uk Chun, J., & Dansereau, F. (2008). Multi-level nature of and multi-level approaches to leadership. *The Leadership Quarterly*, *19*(2), 147–157.

Zhang, Y., Liu, X., & Chen, D. (2023). GIST: Guided in-structure training for contrastive learning. *International Conference on Machine Learning*, 12456–12468.

# Tables

# Figures

**Figure 1**

*t-SNE Visualization of the Five Most Coherent Constructs from IPIP Personality and Leadership Domains*

Note. The left panel displays the five most semantically coherent personality constructs from the held-out IPIP sample (87.4% overall accuracy, Cohen’s d = 1.116), showing tight, well-separated clusters. The right panel shows the five most coherent leadership constructs (62.9% overall accuracy, Cohen’s d = 0.368), displaying more dispersed and overlapping patterns. Each point represents an individual measurement item, with colors indicating construct membership. The visualization uses t-SNE (t-distributed Stochastic Neighbor Embedding) to project high-dimensional semantic embeddings into two dimensions while preserving local structure. The marked difference in clustering patterns provides visual evidence of the differential semantic distinctiveness between personality and leadership measurement domains.

Table 1

Leadership Constructs Included in the Analysis

Leadership constructs included in the semantic analysis with valence, item counts, and example dimensions measured

| Construct | Valence | Number of Items | Example Dimensions |
| --- | --- | --- | --- |
| Transformational Leadership | Positive | 45 | Idealized influence, Inspirational motivation |
| Authentic Leadership | Positive | 38 | Self-awareness, Relational transparency |
| Ethical Leadership | Positive | 32 | Ethical conduct, Moral management |
| Servant Leadership | Positive | 42 | Stewardship, Empowerment |
| Charismatic Leadership | Positive | 28 | Vision articulation, Environmental sensitivity |
| Empowering Leadership | Positive | 26 | Delegation, Coaching |
| Participative Leadership | Positive | 24 | Consultation, Joint decision-making |
| Abusive Supervision | Negative | 35 | Hostile verbal behavior, Public criticism |
| Toxic Leadership | Negative | 42 | Self-promotion, Unpredictability |

Table 2

Comparison of Holdout Validation Approaches

Dual holdout validation framework specifications and rationale

| Approach | Training Data | Holdout Data | Primary Advantage | Validation Focus |
| --- | --- | --- | --- | --- |
| Stratified Item Holdout | 90% of items (2,962) from all 246 constructs | 10% of items (380) from all 246 constructs | Controls for sample size effects | Item-level generalization within known constructs |
| Complete Construct Holdout | All items from 197 constructs (80%) | All items from 50 held-out constructs (20%) | Tests novel construct generalization | Construct-level generalization to novel dimensions |

Table 3

Construct Separation Performance Comparison

Performance metrics comparing construct separation between personality and leadership domains

| Domain | Clustering.Accuracy | X95..CI | Adjusted.Rand.Index | Cohen.s.d | Effect.Size |
| --- | --- | --- | --- | --- | --- |
| IPIP Personality | 87.4% | [85.2, 89.6] | 0.78 | 1.116 | Large |
| Leadership Constructs | 62.9% | [58.7, 67.1] | 0.31 | 0.368 | Small-Medium |
| Performance Difference | 24.5 pp | [18.7, 30.0] | 0.47 | 0.748 | Large |