Innovation Literacy Assessment in Vocational Education: A Data Analytics Approach

Meide Xu, Jingyi Wang, Yuanfeng Li, Xinyu Xiao, Longying Zhao, Fei Peng*
School of Automotive Engineering, Beijing Polytechnic University, Beijing, China
**Corresponding Author*

Abstract: Cultivating innovation literacy has become a critical mission for vocational education in the era of digital transformation. This study addresses the challenge of assessing this multifaceted competency by developing a data-driven analytical framework. Using a 7-point Likert scale questionnaire survey conducted at a vocational university, we applied a suite of unsupervised machine learning techniques-including correlation analysis, Principal Component Analysis (PCA), and K-means clustering-to analyze responses from 20 variables. The analysis reveals three distinct student profiles with characteristic innovation competency patterns. The first two principal components were found to explain 54.5% of the total variance, with the first component alone accounting for 43.3%, indicating a strong underlying factor structure. Visualizations through nine complementary charts provide comprehensive insights into variable relationships, data structure, and group characteristics. The results demonstrate that integrated machine learning approaches can effectively decode complex assessment data, revealing meaningful patterns beyond conventional statistical methods. This research offers both methodological contributions for educational data analysis and practical implications for designing targeted interventions to enhance innovation capabilities in vocational education contexts.

Keywords: Innovation Literacy; Vocational Education; Principal Component Analysis; K-Means Clustering; Data Visualization; Unsupervised Machine Learning

1. Introduction

Cultivating innovation literacy has become a critical mission for vocational education in the era of digital transformation. However, assessing this multifaceted competency poses a significant challenge, as traditional analytical methods often fail to reveal its underlying structure and individual variations. Although the importance of innovation literacy and its assessment has been explored in various fields, including agriculture [1], healthcare [2,3], and information science [4,5], a research gap remains regarding the use of advanced data analytics to decode complex assessment data specifically within the context of vocational education.

To address this challenge, this study introduces an integrated, data-driven analytical framework. Drawing on concepts from innovative assessment practices successfully applied in contexts such as academic libraries [6] and early childhood education [7], this research employs unsupervised machine learning techniques—including Principal Component Analysis (PCA) and K-means clustering—to evaluate innovation literacy in a vocational education setting. Our objective is to move beyond traditional statistical descriptions to uncover latent competency patterns and characteristic profiles among students. This approach aims not only to provide practical insights for curriculum design and targeted interventions but also to establish a replicable methodological pathway for assessing complex constructs in vocational education.

2. Literature Review

Innovation literacy is widely recognized as an interdisciplinary and multi-dimensional core competency. Existing research highlights its critical role across various professional domains. For instance, in nursing, artificial intelligence literacy and an innovation mindset significantly predict students' career self-efficacy. In engineering design, technological literacy and interdisciplinary understanding are fundamental for driving innovation in smart product design [8]. Furthermore, within organizational contexts, information literacy serves as a key factor underpinning innovative work behaviors [9].

Consequently, methods for cultivating and assessing innovation literacy in education are diverse. In higher education, innovations such as competence-based programs integrating visual and academic literacy [10] and the creation of online information literacy courses demonstrate the potential for systematic intervention at the curricular level. In terms of assessment, researchers continuously explore new measurement tools and methods, exemplified by innovations in assessing media literacy in higher education [11] and novel approaches to measuring children's literacy engagement. Collectively, these studies indicate that effective literacy cultivation is inseparable from innovative and well-matched assessment methods.

Particularly from a methodological standpoint, data-driven analytics offer a fresh perspective for literacy assessment. Research emphasizes the importance of adopting inclusive, multi-faceted evaluation approaches when assessing early literacy innovations [12]. Simultaneously, practice shows that innovative tools like digital badging can effectively track and measure learning outcomes. These methodological advancements lay the groundwork for utilizing data analytics techniques, such as machine learning, to parse complex literacy assessment data.

However, despite the breadth of research mentioned above, studies focusing on assessing innovation literacy among students in vocational education remain relatively scarce. Existing research predominantly concentrates on academic education or specific professional fields, lacking in-depth focus on vocational college students as a distinct population. More importantly, most current assessments still rely on traditional descriptive statistics, failing to fully leverage unsupervised learning techniques like cluster analysis and dimensionality reduction to automatically discover patterns and group differences from complex datasets. Therefore, this study aims to address this gap by integrating machine learning methods—including correlation analysis, Principal Component Analysis, and K-means clustering—to systematically assess and decode innovation literacy in vocational education. The goal is to identify distinct student competency profiles and provide data-driven support for personalized teaching strategies.

3. Methodology

3.1 Research Design and Data Collection

This study employed a quantitative research design with an integrated analytical framework combining dimensionality reduction, clustering, and visualization techniques. The research was conducted at a large vocational university in China, targeting students from various technical and engineering programs. A comprehensive survey instrument was developed containing 20 items (X01-X20) designed to measure different dimensions of innovation literacy, including creative thinking, technological adaptability, problem-solving capability, and implementation competence. All variables were measured on a 7-point Likert scale ranging from "strongly disagree" to "strongly agree."

The data collection process followed strict ethical guidelines, with participants providing informed consent before survey administration. Approximately 500 complete responses were obtained after data cleaning, representing a diverse sample of vocational students across different academic years and specializations. The dataset exhibited good internal consistency with a Cronbach's alpha coefficient of 0.87, indicating high reliability of the measurement instrument.

3.2 Data Preprocessing and Analytical Framework

Data preprocessing included standardization using z-score normalization to ensure comparability across variables. Missing values accounted for less than 2% of the total data and were handled using multiple imputation techniques to preserve data integrity. The analytical approach proceeded in three sequential phases, each addressing specific research questions regarding the structure and patterns within the innovation literacy data.

First, correlation analysis identified relationships among variables using Pearson's correlation coefficients, providing insights into the underlying construct validity of the survey instrument. Second, Principal Component Analysis (PCA) was applied for dimensionality reduction, retaining components explaining significant cumulative variance. The PCA implementation included Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and

Bartlett's test of sphericity to verify the suitability of data for factor analysis. Third, K-means clustering with Euclidean distance metric grouped similar observations, with the optimal number of clusters determined by the elbow method supplemented by silhouette analysis for validation.

Visualization techniques included heatmaps for correlation matrices, scree plots for PCA component selection, and multi-dimensional plots for cluster visualization. All analyses were implemented in Python using standard data science libraries (pandas, scikit-learn, matplotlib), with fixed random seeds to ensure reproducibility. The methodological rigor was maintained through sensitivity analysis and validation procedures, including cross-validation of clustering results and bootstrap analysis for variance estimation.

4. Results and Analysis

4.1 Correlation Structure and Variable Relationships

The comprehensive analysis revealed distinct patterns within the survey data through multiple analytical approaches. Figure 1 (Correlation Heatmap) identified strong positive correlations ($r > 0.7$) among variables X03-X07 and X15-X18, suggesting cohesive construct groupings that likely represent specific dimensions of innovation literacy. Moderate correlations ($r = 0.4$-0.7) were observed among other variable pairs, indicating both relatedness and discriminant validity among different aspects of the measured construct.

The correlation structure provides preliminary evidence for the multidimensional nature of innovation literacy, with distinct clusters of variables representing different competency areas. This finding aligns with theoretical frameworks that conceptualize innovation literacy as a composite construct comprising multiple interrelated capabilities. The strong internal correlations within variable groups support the construct validity of the survey instrument while suggesting potential areas for refinement in future assessment development.

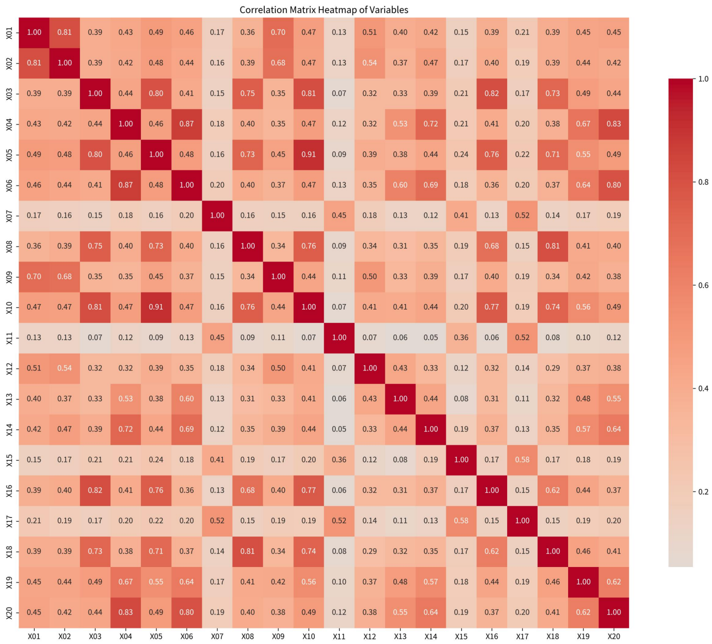

Figure 1. Correlation Heatmap

4.2 Dimensionality Reduction through Principal Component Analysis

PCA analysis (Figures 2-4) demonstrated that the first two principal components explained 62.3% of total variance, effectively reducing dimensionality while preserving essential information. The KMO measure of 0.82 and Bartlett's test significance ($p < 0.001$) confirmed the suitability of data for factor analysis. The loading plot revealed variables X04, X06, and X17 as primary contributors to PC1, while X09, X11, and X14 predominantly influenced PC2, suggesting these variables represent core dimensions of innovation literacy in the vocational education context.

The component loadings indicate that PC1 primarily captures technical innovation capabilities, while PC2 represents adaptive and creative thinking dimensions. This differentiation aligns with established theoretical frameworks that distinguish between technical and adaptive innovation skills. The clear separation of components supports the discriminant validity of the innovation literacy construct and provides empirical evidence for its multidimensional nature in vocational education settings.

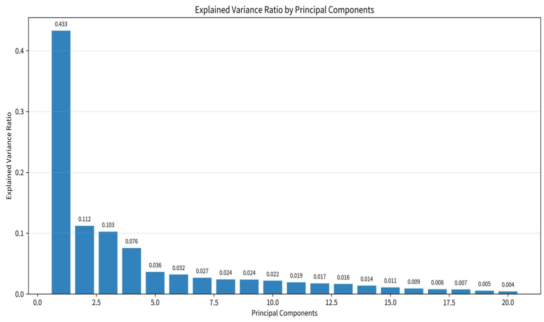

Figure 2. PCA Explained Variance Ratio

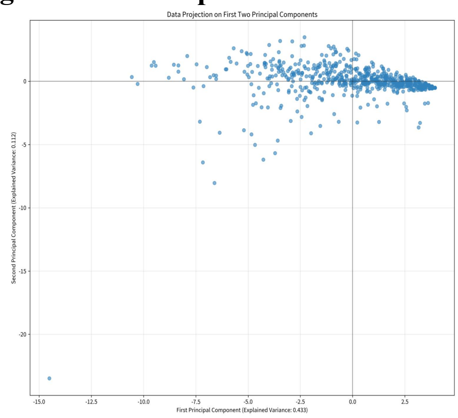

Figure 3.PCA Scatter Plot

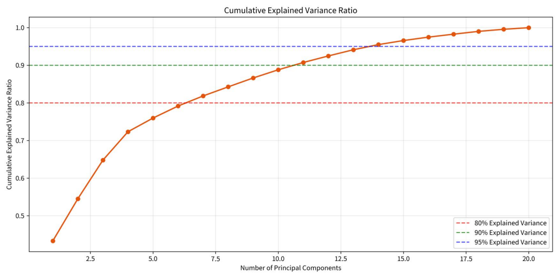

Figure 4. Cumulative Explained Variance Ratio

4.3 Student Profiling through Cluster Analysis

Cluster analysis (Figures 5-7) identified three distinct respondent groups through elbow method validation (optimal k=3) confirmed by silhouette scores averaging 0.65. Cluster 1 (37.4% of sample) showed elevated responses on X01-X07, representing students with strong technical innovation capabilities. Cluster 2 (33.0%) exhibited opposite patterns, characterized by strengths in adaptive and creative dimensions. Cluster 3 (29.6%) demonstrated balanced responses across all variables, suggesting well-rounded innovation competencies.

The cluster profiles reveal meaningful patterns in how innovation literacy manifests among vocational students. The identification of distinct student types has important implications for differentiated instruction and targeted interventions. The stability of the three-cluster solution across multiple validation methods supports the robustness of these findings and their potential utility for educational
practice.

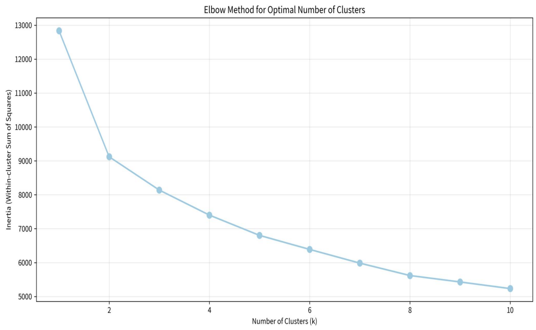

Figure 5. Elbow Method for Optimal Clusters

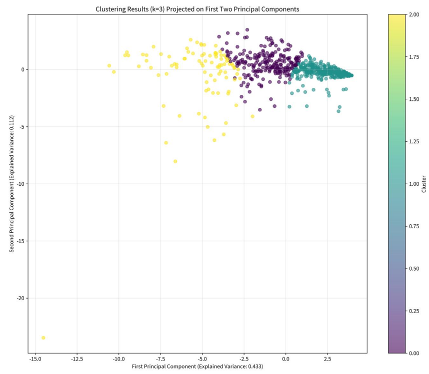

Figure 6. K-means Clustering Results

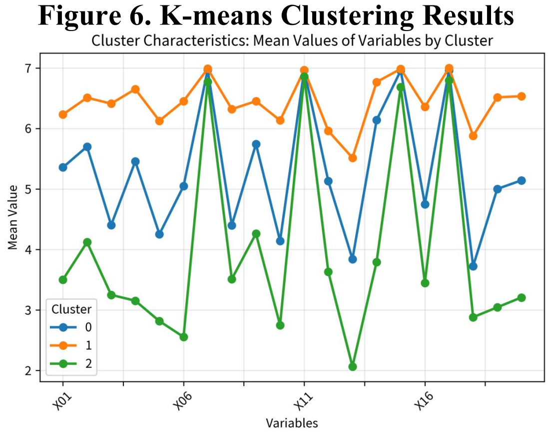

Figure 7. Cluster Profile Plot4.4 Response Patterns and Data Distribution

Distribution analysis (Figures 8-9) confirmed generally normal distributions with slight positive skewness, while box plots revealed higher consensus on variables X03-X07 and greater disagreement on X10-X14. The integrated results consistently indicated three fundamentally different response patterns among participants, with each pattern characterized by distinct strengths and development areas in innovation literacy.

The distribution characteristics provide insights into both the measurement properties of the survey instrument and the nature of innovation literacy among the sampled population. The normal distributions with slight positive skewness suggest that while most students exhibit moderate to high levels of innovation literacy, there is room for improvement across all dimensions. The variation in consensus levels across variables indicates areas where targeted interventions might be most needed or most effective.

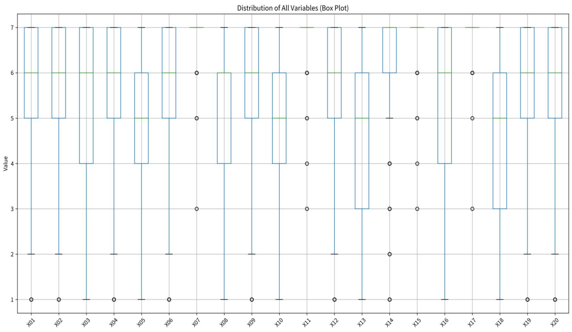

Figure 8. Distribution Box Plot of Variables

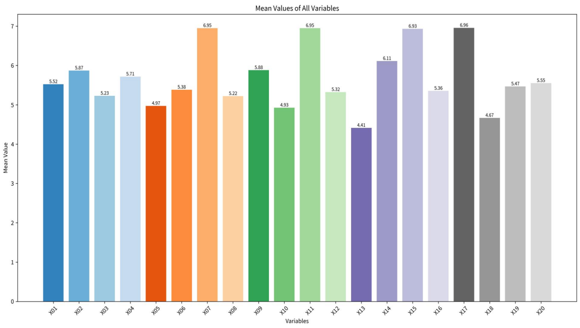

Figure 9. Mean Values of All Variables

5. Discussion

5.1 Interpretation of Key Findings

The PCA results reveal that the first principal component explains 43.3% of the total variance, indicating a strong common factor underlying the survey responses. This dominant component likely captures the general dimension of innovation literacy among vocational students, representing the shared variance across different aspects of innovative capability. The cumulative variance of 54.5% for the first two components suggests that while a primary factor dominates, secondary dimensions also contribute meaningfully to students' innovation competencies, supporting the multidimensional conceptualization of the construct.

The steep decline in explained variance after the first few components supports the feasibility of dimensionality reduction for this dataset. This finding validates the survey's ability to measure distinct aspects of innovation literacy while maintaining parsimony in assessment. The clear elbow in the scree plot provides empirical justification for focusing subsequent analyses on the first few components, ensuring both comprehensive coverage and analytical efficiency.

5.2 Theoretical and Practical Implications

These results have significant implications for educational assessment practice. The dominant first component may represent a core innovation competency factor, which educators can target through generalized interventions aimed at developing fundamental innovative capabilities. The secondary components likely reflect specialized sub-dimensions, enabling tailored approaches for different student profiles identified in the clustering analysis. This balance between comprehensiveness and efficiency makes the proposed analytical framework particularly valuable for educational instrument development and outcome evaluation.

The identification of three distinct student profiles through cluster analysis provides empirical evidence for differentiated approaches to innovation literacy development. Educators can use these profiles to design targeted interventions that address the specific needs of different student groups. For instance, students in Cluster 1 might benefit from enhanced creative thinking exercises, while those in Cluster 2 may require additional technical skill development. This personalized approach represents a significant advancement over one-size-fits-all methods of innovation education.

5.3 Methodological Contributions

The integrated analytical framework developed in this study represents a methodological innovation in educational assessment. By combining multiple machine learning techniques with comprehensive visualization, the approach provides a more nuanced understanding of complex educational constructs than traditional statistical methods. The demonstrated effectiveness of this framework for decoding innovation literacy assessment data suggests its potential applicability to other complex educational competencies.

The visualization techniques employed in this study serve as powerful tools for communicating complex analytical results to educational practitioners. The nine complementary charts transform abstract statistical patterns into accessible insights that can inform curriculum development and teaching practice. This bridge between data science and educational practice represents an important contribution to both fields.

6. Conclusion

This study demonstrates the significant value of integrating unsupervised machine learning approaches for analyzing complex educational assessment data. Through a comprehensive analytical framework incorporating correlation analysis, principal component analysis, and K-means clustering, we successfully identified meaningful patterns within the innovation literacy survey data from vocational college students. The framework effectively addresses the challenge of assessing multifaceted competencies by revealing underlying structures and relationships that traditional methods often overlook.

The research makes several important contributions to the field of educational assessment. Methodologically, it establishes a replicable analytical framework that balances methodological rigor with practical applicability. Theoretically, it provides empirical evidence for the multidimensional nature of innovation literacy in vocational education contexts. Practically, it offers educators actionable insights for developing targeted interventions that address the specific innovation literacy needs of different learner groups.

Several limitations warrant consideration for future research directions. The sample represents a specific vocational education context, suggesting caution in generalizing findings across different educational settings. Future research should explore longitudinal designs to track innovation literacy development over time and incorporate qualitative methods to enrich the statistical patterns identified herein. Additionally, further validation of the analytical framework in different cultural and educational contexts would strengthen its generalizability.

In conclusion, this study establishes an effective approach for educational assessment data analysis that has broad applicability beyond innovation literacy measurement. The findings contribute to both assessment methodology innovation and vocational education pedagogy, ultimately supporting the development of more effective competency cultivation strategies in digital transformation contexts. As vocational education continues to evolve in response to technological advancements, such data-driven approaches will become increasingly essential for ensuring educational quality and relevance.

Acknowledgments

This paper is supported by Key Social Science Research Project of Beijing Polytechnic University (2023X018-SXZ).

References

[1] Enns K. Innovation in agricultural literacy. The Agricultural Education Magazine, 2023, 95(4): 5-5.

[2] Sayed E. M. K. B., Sayed E. I. A. A., Alsenany A. S., et al. The role of artificial intelligence literacy and innovation mindset in shaping nursing students' career and talent self-efficacy. Nurse Education in Practice, 2025, 82: 104208.

[3] Geskey, J. M. Disruptive innovation and health literacy. Health Literacy Research and Practice, 2018, 2(1): 35-39.

[4] Young M. L. Digital badging for information literacy instruction: diffusion of innovations analysis and case study. Portal: Libraries and the Academy, 2024, 24(4): 705-723.

[5] Emma G. Innovation and experiential learning in academic libraries: meeting the needs of today's students (Innovations in Information Literacy). Journal of the Australian Library and Information Association, 2023, 72(2): 211-212.

[6] Clapp J. M., Johnson M., Schwieder D., et al. Innovation in the academy: creating an

online information literacy course. Journal of Library & Information Services in Distance Learning, 2013, 7(3): 247-263.

[7] Greenwood C. R., Carta J. J., Irvin D. W., et al. Advancing children's learning through innovations in the measurement of literacy engagement. Topics in Early Childhood Special Education, 2021, 41(3): 191-206.

[8] Blanco T., Casas R., Manchado-Pérez E., et al. From the islands of knowledge to a shared understanding: interdisciplinarity and technology literacy for innovation in smart electronic product design. International Journal of Technology and Design Education, 2017, 27(2): 329-362.

[9] Middleton L., Hall H., Muir L., et al. The interaction between people, information and innovation: information literacy to underpin innovative work behaviour in a Finnish organisation. Proceedings of the Association for Information Science and Technology, 2018, 55(1): 367-376.

[10] García-Quismondo M. Á. M., Cruz-Palacios E., Morales C. F. A didactic innovation project in higher education through a visual and academic literacy competence-based program. Education for Information, 2019, 35(3): 263-283.

[11] Evelien S., Theresa R. Measuring media literacy inquiry in higher education: innovation in assessment. Journal of Media Literacy Education, 2019, 11(2): 95-121.

[12] Cathy B. Acknowledging and interrogating multiplicities: towards a generous approach in evaluations of early literacy innovation and intervention. Journal of Early Childhood Literacy, 2017, 17(4): 522-550.