STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–28.
Published online in International Academic Press (www.IAPress.org)

Exploring Model-Dependent Learning Profiles in PIRLS 2021: A Comparative Study of Classical and Deep Clustering Methods

Rim Elmaataoui [1,*], Soukaina Elharouri [2], Soufiane Lyaqini [1], Imad Hafidi [1]

[1] *IDIA Laboratory, National School of Applied Sciences, Sultan Moulay Slimane University, Khouribga, Morocco*
[2] *LAMDA, National School of Applied Sciences of Khouribga, Sultan Moulay Slimane University, Morocco.*

Abstract Large-scale educational assessments combine heterogeneous information that complicates learner profiling, while deep clustering can produce apparently well-separated latent groups whose educational relevance remains uncertain. Using 7,017 Moroccan fourth-grade student records from PIRLS 2021, this study compares classical clustering methods (KMeans, agglomerative clustering, and Gaussian mixture models) with PCA+KMeans and three deep approaches: AE+KMeans, Deep Embedded Clustering (DEC), and Improved DEC (IDEC). To ensure a fair comparison, all partitions are evaluated in the same standardized input space, while latent-space indices are treated as representation-specific diagnostics. We further assess assignment stability across random seeds, sensitivity to preprocessing choices, and associations with held-out reading achievement, home learning resources, and home-language frequency. DEC and IDEC achieve high latent-space Silhouette scores (0.9994 and 0.8845, respectively), but their common-space scores remain low (0.0221 and 0.0245). The resulting profiles exhibit modest associations with reading achievement (weighted $\eta^2 = 0.049$ for DEC and 0.045 for IDEC) and stronger associations with home resources (0.078 and 0.067), whereas home-language associations remain weak. Assignment stability is low (mean adjusted Rand index = 0.221 for DEC and 0.222 for IDEC), and individual memberships are sensitive to missing-data treatment and the inclusion of an identifier-like variable. A matched 16-component PCA+KMeans baseline is substantially more stable (mean ARI = 0.702) and outperforms the deep methods on all three common-space validity indices. Harmonized replications in France, Singapore, and South Africa confirm the discrepancy between latent and common-space separation, although educational effects vary across countries. The computational benchmark also reveals substantially higher costs for deep methods. Overall, these findings support the cautious use of deep clustering for exploratory educational profiling but do not establish stable learner categories or a clear practical advantage over simpler alternatives.

Keywords Deep clustering, representation learning, k-means, agglomerative clustering, GMM, PIRLS 2021, reading achievement, unsupervised learning, learning analytics.

AMS 2010 subject classifications 62H30, 68T07

DOI: 10.19139/soic-2310-5070-4191

1. Introduction

Clustering is a central tool in unsupervised learning, widely used to explore hidden patterns in data when no predefined labels are available. Traditional methods such as *k*-means, agglomerative clustering, and Gaussian Mixture Models (GMM) remain popular because they are easy to implement and interpret. Yet, their effectiveness tends to decline when applied to high-dimensional, noisy, or heterogeneous datasets situations in which distance-based similarity becomes unreliable and the underlying structure is often non-linear.

*Correspondence to: Rim Elmaataoui (Email: rim.elmaataoui@usms.ac.ma). IDIA Laboratory, National School of Applied Sciences, Sultan Moulay Slimane University, Khouribga, Morocco.

ISSN 2310-5070 (online) ISSN 2311-004X (print)
Copyright © 202x International Academic Press

Across different application domains, the strengths and limitations of classical clustering have become increasingly evident. In text mining, for instance, clustering has been employed for document organization, corpus structuring, and the reduction of lexical redundancy [1, 2]. Similar methodological trends are observable in the Moroccan research landscape, where machine learning has been used to analyze sentiment in Moroccan Darija [3], infer personality traits from online behavior [4], develop linguistic resources [5], and forecast economic indicators using modern predictive models [6]. These studies reflect a growing interest in data-driven approaches for understanding complex, multidimensional phenomena.

In education, large-scale international assessments such as TIMSS and PIRLS offer rich multilevel datasets that capture both learning outcomes and contextual variables. A recent review of machine-learning applications and datasets in education also emphasizes the importance of data quality, appropriate evaluation, and responsible interpretation in this domain [7]. Previous analyses have examined mathematics achievement using TIMSS data [8] and explored determinants of reading performance in Morocco [9]. Other work has investigated learners' online behavior through clickstream-based prediction models [10]. While these contributions highlight the promise of machine learning for educational research, they also expose the limitations of shallow clustering techniques when dealing with heterogeneous educational data.

Related work on complex survey data has combined model-based and distance-based clustering to identify child vulnerability profiles, while reporting limited agreement between the individual assignments produced by the two methods [11]. This result further motivates assessing assignment stability rather than interpreting a single partition as a definitive learner taxonomy.

Recent progress in representation learning has given rise to deep clustering methods, which jointly learn feature representations and clustering assignments within a unified framework. By leveraging neural networks to construct latent spaces that reduce noise and capture non-linear relations, these methods have shown strong results in fields such as computer vision, natural language processing, and bioinformatics. Their ability to reshape the data geometry makes them particularly appealing for complex datasets where traditional clustering may fail to uncover meaningful structure.

In this study, we explore the use of deep clustering for the **PIRLS 2021 International Database**, a comprehensive assessment of reading comprehension administered to fourth-grade students in more than fifty education systems [12]. The Moroccan subset combines achievement data with extensive contextual information from students, teachers, and schools, creating a challenging setting for unsupervised analysis. Its heterogeneous nature provides an opportunity to assess how strongly the resulting profiles depend on the learned representation and whether they have educational utility beyond favorable latent-space geometry.

To investigate this question, we compare several deep clustering techniques with standard approaches such as k-means, agglomerative clustering, and GMM. We also include PCA+KMeans as a matched non-neural dimensionality-reduction baseline. We distinguish internal validity in the optimized latent space from validity in the common original feature space, examine assignment stability across random seeds, and evaluate whether held-out reading achievement and educational variables differ across the resulting profiles.

Contributions

This work offers five main contributions:

1. It applies deep clustering as an exploratory representation-learning framework to the Moroccan PIRLS 2021 data without assuming that the resulting partitions are natural or ground-truth learner categories.
2. It provides a fair cross-space comparison between classical, PCA-based, and deep clustering. The results show that strong separation in an optimized latent representation need not be preserved when the same labels are evaluated in the original standardized space, and that the present data do not establish an advantage of complex deep objectives over linear compression followed by KMeans.
3. It evaluates the profiles beyond internal metrics through held-out reading plausible values, external educational variables, and assignment-level stability across random initializations. Missing-data sensitivity and hyperparameter ablation further identify which conclusions and memberships depend on preprocessing and representational capacity.

4. Rather than presenting established contextual correlates as new discoveries, it examines how they combine within multivariate profiles and shows that the resulting configurations are overlapping, method-dependent, and not reducible to a single language or socioeconomic hierarchy.
5. It examines the scope and practical cost of the framework through a controlled computational benchmark and harmonized replications in France, Singapore, and South Africa. This distinguishes a cross-country robust methodological warning from country-specific profile composition and educational effect sizes, while documenting the additional time and memory required by the deep methods.

2. Related work

2.1. Classical Clustering Methods

The process of clustering is an unsupervised machine learning process, which means that the goal of clustering is to divide the data into groups on the basis of their similarity, without any specific labels. For the purpose of this study, conventional clustering algorithms were taken into consideration to compare their performance with that of the deep clustering models.

1. **k-means** [13, 14] was used as a centroid-based clustering algorithm in this study. The k-means algorithm divides the data into a number of groups, where the number of groups is determined by the number of centroids. It works by iteratively assigning each data point to the nearest centroid, followed by updating the centroid positions. It is a commonly used baseline algorithm due to its simplicity and ease of use. It is, however, better suited for cases in which the data points are well separated.
2. **Agglomerative clustering** [15] was used as a hierarchical clustering algorithm in this study. Agglomerative clustering starts with each data point as a separate cluster, which are then merged to form larger clusters using a linkage function. It can handle complex structures in the data, but the quality of the results relies on the linkage function used for the purpose of clustering.
3. **Gaussian Mixture Model (GMM)** [16] was used as a probabilistic clustering algorithm in this study. GMM is a type of clustering algorithm that can handle complex structures in the data by considering the data to be distributed in accordance with a mixture of Gaussian distributions. The quality of the clustering process using GMM, however, relies on the quality of the assumption of the data to be distributed in accordance with a Gaussian distribution.

These techniques were used in our experiments as traditional baselines, whereas the specific implementation details of each method are discussed in the experimental setup section.

2.2. Standard Dimensionality-Reduction Baseline

To separate the contribution of dimensionality reduction from that of a clustering-specific neural objective, we added PCA followed by k-means as a non-neural representation-learning baseline. PCA constructs orthogonal linear components ordered by explained variance, after which k-means is applied to the retained scores. Unlike AE+KMeans, DEC, and IDEC, this procedure has neither a non-linear encoder nor a reconstruction or KL-divergence objective. It therefore tests whether representation compression alone can account for the apparent advantage of latent-space clustering. The PCA representation was fixed at 16 components to match the dimensionality of the autoencoder bottleneck, and the resulting labels were also evaluated in the common standardized input space.

2.3. Deep Clustering

Deep clustering [17] is a type of clustering that seeks to unify representation learning with cluster assignment in a single framework. Instead of using a clustering algorithm on the original variables, deep clustering methods use a neural network to learn a representation of the data, followed by a clustering algorithm on the learned

representation. The intuition behind this type of clustering is that the learned representation can better capture the internal structure of the data than the original variables.

When compared to the general clustering algorithm, this type of clustering can be used in scenarios where the data contains complex relationships, noisy variables, or non-linear relationships. By mapping the original variables to a new space, which is generally lower-dimensional, compact, and informative, the separation of the clusters can be improved, making them easier to interpret.

This type of clustering has been used in several scenarios with high-dimensional, complex, and structured data. For this paper, the use of this type of clustering can be justified because, in the PIRLS 2021 dataset, there is a mixture of variables from students, teachers, and schools. This type of data can be complex to structure using general clustering algorithms on the original variables. Thus, using deep clustering to learn a new representation of the data, which can better capture the internal structures of the data, was deemed appropriate for this dataset.

2.3.1. Deep Embedded Clustering (DEC) Deep Embedded Clustering (DEC) [18] is another deep clustering algorithm that aims at jointly learning the representation and refining the clustering assignments over time. This process usually starts with the autoencoder, which is then utilized to project the input data onto the feature space with reduced dimensionality. After the initialization process is done, the clustering is performed directly in the reduced space using the centroids.

The main concept behind the DEC algorithm is the refinement of the clustering assignments during training. To achieve this, the model compares the current soft assignments with an auxiliary target distribution that gives greater importance to high-confidence assignments. This mechanism gradually encourages the latent space to become more suitable for clustering.

Given a latent representation z_n and a cluster centroid μ_c, the soft assignment is defined using a Student's t-distribution kernel:

$$q_{nc} = \frac{(1 + \|z_n - \mu_c\|^2/\alpha)^{-(\alpha+1)/2}}{\sum_{c'}(1 + \|z_n - \mu_{c'}\|^2/\alpha)^{-(\alpha+1)/2}}.$$

The auxiliary target distribution is then expressed as:

$$p_{nc} = \frac{q_{nc}^2/f_c}{\sum_{c'} q_{nc'}^2/f_{c'}}, \quad \text{with } f_c = \sum_n q_{nc}.$$

The training objective consists of minimizing the Kullback-Leibler divergence between the target distribution P and the predicted assignment distribution Q:

$$L = KL(P \parallel Q) = \sum_n \sum_c p_{nc} \log \frac{p_{nc}}{q_{nc}}.$$

Through this optimization process, DEC progressively sharpens the cluster structure in the latent space while updating the encoder so that the learned representation becomes increasingly aligned with the clustering objective. In our study, DEC was adopted as one of the main deep clustering baselines to examine the latent structure underlying the PIRLS 2021 data.

2.3.2. Improved Deep Embedded Clustering (IDEC) Improved Deep Embedded Clustering (IDEC) [19] was proposed as an extension of DEC in order to preserve the reconstruction objective during the clustering phase. Unlike DEC, which mainly concentrates on improving cluster assignments in the latent space, IDEC keeps the autoencoder active throughout the optimization process. This allows the learned representation to remain closely related to the original input data.

The main idea behind IDEC is to avoid excessive distortion of the latent space when training is guided only by the clustering objective. To address this issue, the method combines the clustering loss with a reconstruction loss, leading to the following objective function:

$$L_{\text{IDEC}} = L_{\text{c}} + \gamma L_{\text{r}},$$

where the clustering term is the same KL-divergence objective used by DEC,

$$L_{\mathrm{c}} = \mathrm{KL}(P \parallel Q) = \sum_{n=1}^{N} \sum_{c=1}^{K} p_{nc} \log\left(\frac{p_{nc}}{q_{nc}}\right),$$

and the reconstruction term is

$$L_{\mathrm{r}} = \frac{1}{N} \sum_{n=1}^{N} \|g_\phi(f_\theta(\mathbf{x}_n)) - \mathbf{x}_n\|_2^2,$$

with encoder f_θ, decoder g_ϕ, and reconstruction $\hat{\mathbf{x}}_n = g_\phi(f_\theta(\mathbf{x}_n))$. Thus, L_{r} is exactly the mean-squared reconstruction criterion used to pretrain the autoencoder and used by AE+KMeans below; the difference is that IDEC retains this term jointly with L_{c} during clustering refinement. In the primary experiments, $\gamma = 1$.

By optimizing these two components jointly, IDEC aims to balance cluster compactness with preservation of the input structure. In this study, IDEC was included to test whether retaining the reconstruction objective reduces representation distortion relative to DEC; stability is evaluated empirically rather than assumed from the objective.

2.3.3. Autoencoder with K-Means (AE+KMeans) Autoencoder with K-Means (AE+KMeans) is a straightforward deep clustering baseline in which representation learning and clustering are carried out in two distinct steps. In the first stage, an autoencoder is trained to reconstruct the input data while learning a lower-dimensional latent representation. In the second stage, K-Means is applied to this latent space rather than to the original feature space.

The reconstruction objective of the autoencoder is defined as:

$$L_r = \frac{1}{N} \sum_{n=1}^{N} \|\hat{\mathbf{x}}_n - \mathbf{x}_n\|_2^2.$$

Once training is completed, the encoder maps each sample $\mathbf{x}_n$ to a latent vector $\mathbf{z}_n$, and clustering is then performed on these latent embeddings. Unlike DEC and IDEC, AE+KMeans does not update the latent space through a clustering-specific objective during training. Its purpose is therefore to evaluate whether a learned representation, followed by a standard clustering algorithm, is sufficient to uncover a meaningful structure in the data.

In our study, this method was considered an intermediate baseline between purely classical clustering techniques and deep clustering approaches in which representation learning and cluster optimization are performed jointly.

3. Problem Statement

Large-scale educational assessments combine continuous, ordinal, and categorical information from students, teachers, and schools. Their high dimensionality, correlated indicators, missing values, and heterogeneous scales complicate any distance-based analysis. Classical methods such as *k*-means, agglomerative clustering, and GMM remain transparent and computationally accessible, but their partitions can depend on scaling, distance geometry, initialization, linkage, or covariance specification. Representation learning offers a possible way to summarize this information before clustering, but it introduces a different methodological risk: the learned space can be reshaped by the training objective itself. This risk is especially important for DEC and IDEC. Their clustering objectives can move embeddings toward centroids and sharpen assignments, so high Silhouette or Calinski-Harabasz values in the optimized latent space do not necessarily imply that the same groups are well separated in the observed educational variables. Directly comparing a deep method in its learned space with a classical method in the original space therefore confounds the clustering procedure with the feature geometry. The central problem is not simply to maximize latent compactness, but to determine whether the resulting partitions represent stable and educationally informative profiles rather than model-induced divisions of continuous gradients. Accordingly, this study asks whether AE+KMeans, DEC, and IDEC provide evidence beyond classical clustering and a matched PCA+KMeans baseline when all labels are evaluated in the same standardized input space. It also examines whether

the assignments are stable across random initializations, missing-data treatments, and hyperparameter choices; whether variables excluded from clustering, including reading achievement, home resources, and home-language frequency, differ meaningfully across profiles; whether the principal findings recur in other PIRLS countries; and whether any added value justifies the computational cost of the deep models. In the absence of ground-truth classes, these complementary criteria are required before the clusters can be interpreted as useful learner profiles.

4. Data Description

The dataset used in this study was obtained from the PIRLS 2021 International Database. The database was constructed by the International Association for the Evaluation of Educational Achievement and is publicly accessible at https://pirls2021.org/data/. The database includes data collected from fourth-grade students and their teachers, parents, and school principals in more than fifty education systems. PIRLS 2021 marks the beginning of the hybrid PIRLS cycle with the integration of the traditional paper-based assessment and the digital version of the assessment tool, ePIRLS [12, 20].

The international database includes records from approximately 400,000 students, 380,000 parents or caregivers, 20,000 teachers, and 13,000 school principals across 57 countries and 8 benchmarking systems. Each country's data was collected with the help of the IEA's two-stage stratified sampling approach.

Aside from the entire international database of PIRLS 2021, this study specifically utilizes the nationally representative Moroccan subsample of the database. The Moroccan database maintains the entire structure of PIRLS contextual information, including the identifiers and background information required to link and analyze data across the student, teacher, and school levels.

National representativeness should not be confused with international generalizability. Morocco remains the primary analytic sample and the basis of the detailed profile interpretation. To test whether the methodological conclusions depend on this single context, we added complete PIRLS 2021 samples from France, Singapore, and South Africa. These systems provide different sample sizes and contextual, linguistic, and resource distributions. The cross-country analysis tests reproducibility of the cross-space pattern; it does not assume that identically numbered clusters have the same substantive meaning across countries.

The Moroccan subsample of the PIRLS database used in this study consists of:

- **7,017 fourth-grade student records** that include reading achievement information and a wide range of contextual attributes, such as demographics, home environment, and learning behaviors.
- **290 teacher-file records, corresponding to 280 distinct teachers across 266 schools,** that include information on teaching methods, professional backgrounds, resource availability, and learning climate. Teacher variables were subsequently aggregated at the school level.
- **266 school principal questionnaires** that include information on school characteristics, school leadership, school infrastructure, and resources.

Source variable names and coding conventions follow the official PIRLS 2021 structure and were preserved in the audit trail. However, the analytical matrix was subsequently filtered, aggregated, imputed, and standardized as described in the preprocessing section. The following key indicators are included and will be used in later descriptive and graphical analysis:

For the international replication, the analysis retained the intersection of variables that were available, nonconstant, and below the 50% missingness threshold in all four systems. The resulting harmonized matrix contains 149 variables: 15 continuous and 134 discrete variables. Sample sizes are 7,017 for Morocco, 5,339 for France, 6,719 for Singapore, and 12,422 for South Africa. Technical compatibility permits the same pipeline to be fitted in every country, while country-specific standardization avoids interpreting differences in national response levels as geometric separation.

4.1. Students background

The student files contain identifiers, assessment-design variables, and background descriptors. Examples include:

- IDSCHOOL: unique identifier for the school attended by the student.
- IDCLASS: identifier for the classroom within the school.
- IDSTUD: unique identifier for each student in the PIRLS dataset.
- ITSEX: gender of the students. (1.0: Female, 2.0: Male).
- ITLANG_SA: language of the student achievement test. In the Moroccan sample this variable equals 53 (Arabic) for all 7,017 students and is therefore removed by the constant-variable filter. It is distinct from ASBG03, the student-context question on how often the test language is spoken at home.

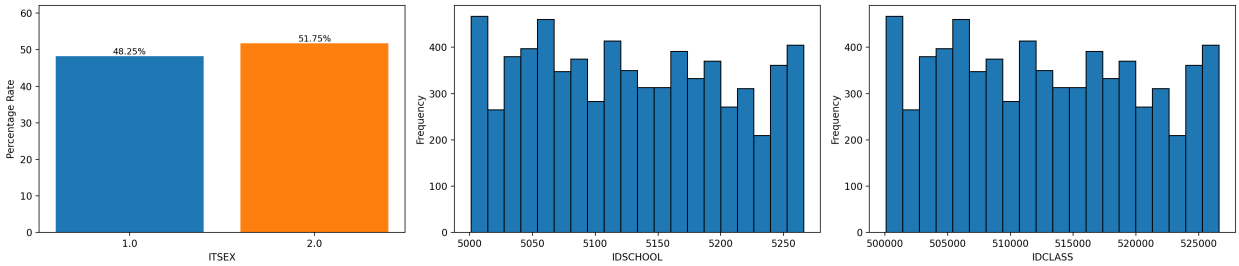


Figure 1. Distributions of Student Gender and School/Class Identifiers in PIRLS 2021

4.2. Teacher background

The dataset for teachers offers information on teachers' profiles, teaching conditions, and professional characteristics. The parameters for this dataset include:

- ATBG01: This parameter is used to describe the gender distribution of teachers in the classroom. The values are: (1.0 Female, 2.0 Male).
- ATBG02: This parameter is used to understand the age group distribution of teachers in the classroom.
- ATBG03: This parameter is used to understand the educational qualifications of teachers in the classroom.
- ATBG04: This parameter is used to understand the years of experience for teachers in the classroom.
- ATBG05AA,ATBG05AE: This parameter is used to understand teachers' participation in professional development activities in the classroom.
- ATBG07A,ATBG07B: This parameter is used to understand teachers' resources and support in the classroom.
- ATBG08A,ATBG08E: These parameters are used to understand teacher-reported classroom practices and instructional strategies.

These teacher-level factors offer important information about the instructional environment, teacher qualifications, and the extent of pedagogical support in the school context.

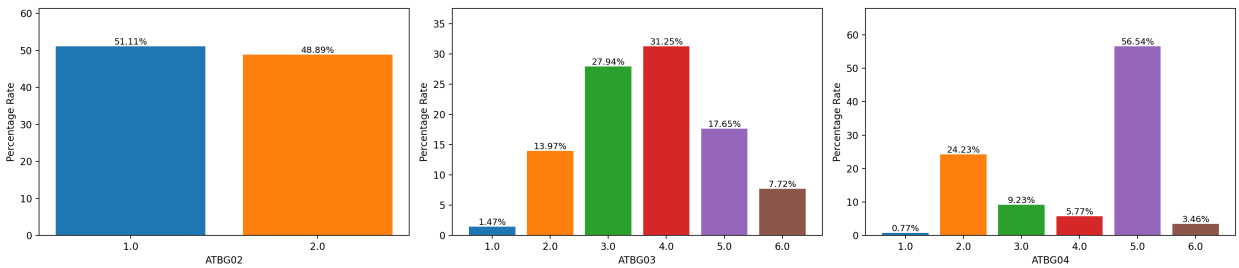


Figure 2. Distributions of Teacher Age, Educational Qualifications, and Teaching Experience in PIRLS 2021.

4.3. School background

The school dataset contains detailed background information on school-level variables, infrastructure, and learning environments. Some of the key parameters include:

- `ACBG03A`, `ACBG03B`: School location, which could include rural, suburban, or urban settings.
- `ACBG04`: School size, which is often used to denote the number of students enrolled in school.
- `ACBG05A`, `ACBG05B`: Availability of school facilities.
- `ACBG06A`, `ACBG06C`: School safety, which could include aspects of student behavior, school climate, or discipline concerns.
- `ACBG07A`: Instructional time or school schedule.
- `ACBG08A`, `ACBG08C`: Access to support services, which could include counseling, guidance, or psychological services.
- `ACBG09A`, `ACBG09D`: Infrastructure or facilities in school, which could include libraries, computer facilities, or learning facilities.

Such school-level indicators give us information on school conditions, resources, school management, and community engagement, allowing for multilevel analysis of the structural aspects that affect student outcomes.

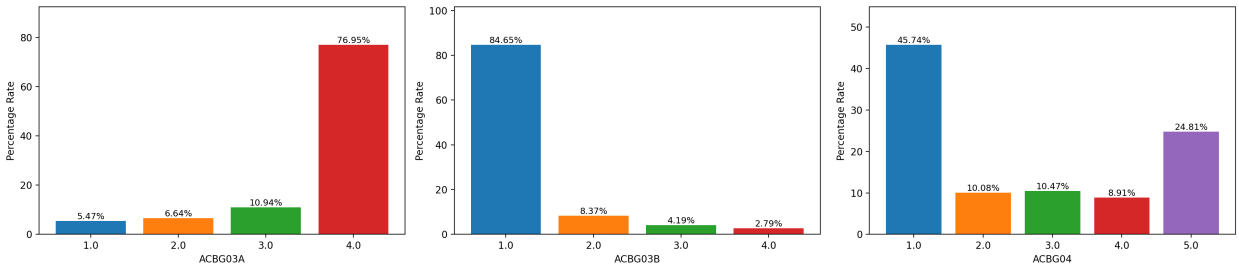


Figure 3. Distributions of School Location and Size Characteristics in PIRLS 2021.

5. Proposed Method

All clustering approaches were fitted from consistently prepared input matrices and evaluated through the method, validation, sensitivity, computational, and cross-country procedures described in the following subsections.

5.1. Data Preprocessing

The PIRLS 2021 data combine continuous, ordinal, and categorical variables from the student achievement/background file and the teacher and school context files. The exported-file missing codes `−99`, `−98`, `−97`, `−96`, `−95`, `999`, and `9999` were converted to `NaN`.

The legacy candidate matrix contained 419 variables, including 193 entirely empty columns. In the revised robust pipeline, missingness and constant-variable filtering were applied within each source block before the source-specific caps. This produced 204 candidate features: 2 student, 120 teacher, and 82 school variables. One additional teacher variable became constant after aggregation and merging, resulting in the final 203-feature matrix: 2 student, 119 teacher, and 82 school variables.

Low-cardinality integer-coded variables were treated as discrete. Missing values in discrete student variables were replaced by the modal observed category, while discrete teacher and school variables were aggregated at school level using their mode. Continuous variables were median-imputed, and continuous teacher and school variables were aggregated using their school-level mean. The resulting matrix contained 181 discrete and 22 continuous features. All features were then standardized to zero mean and unit variance, and the same completed matrix was supplied to every clustering method.

Categorical PIRLS variables were not expanded through one-hot encoding, and no learned categorical embeddings were used. Instead, their retained numerical PIRLS codes were standardized and supplied directly to the models. This compact encoding can impose artificial metric distances between nominal categories, so it is an explicit limitation of the present implementation.

The implementation excluded IDCNTRY, IDSCHOOL, and IDSTUD from the student block but retained IDCLASS. Because IDCLASS is an identifier-like numerical field rather than a substantive educational measure, its inclusion can introduce arbitrary proximity into a distance-based analysis. We report this implementation detail transparently. The archived primary tables retain the matrix used in the reported analysis, while Section 7.14 presents a complete sensitivity analysis after removing IDCLASS.

5.2. Missing-data robustness specifications

To quantify dependence on missing-data handling, we compared the primary matrix with the former median-only specification, which contained 205 effective variables after empty and constant columns were removed. We also generated five stochastic completed datasets using iterative chained-equation-style imputation. Imputed discrete values were mapped back to the nearest category observed for the corresponding variable. Related clustering work has likewise integrated missing-data treatment with robust PCA-based representation construction, illustrating that preprocessing and dimensionality reduction form part of the analytical pipeline [21]. These datasets are used to assess the stability of cluster assignments under imputation uncertainty; they are not treated as five directly observed datasets, nor is this clustering comparison presented as Rubin-rule pooling for a supervised estimand. Partitions were compared with the adjusted Rand index (ARI), which is invariant to arbitrary cluster numbering. The resulting assignment-sensitivity comparisons are reported in Table 7.

5.3. Variables used for clustering and post-hoc interpretation.

The clustering input was constructed from processed variables extracted from the student achievement, teacher, and school PIRLS files. The retained student-file variables are IDCLASS and ITSEX; as noted above, IDCLASS is identifier-like and its inclusion is a limitation. Retained teacher-level variables include ATBG01, ATBG02, ATBG03, and ATBG04, and school-level variables such as ACBG03A, ACBG04, and ACBG05A. Teacher- and school-level information was aggregated at school level and merged with the student records before standardization. Crucially, the five PIRLS overall-reading plausible values (ASRREA01-ASRREA05) were absent from the clustering matrix and were not used by any classical or deep algorithm. They were linked only after cluster estimation and used as held-out outcomes in the external educational validation. Thus, no reading plausible value contributed to the construction of the student profiles. This distinction follows the PIRLS technical documentation, which treats plausible values as multiple draws used to represent achievement uncertainty rather than as five directly observed test scores [20].

The home-language variable used for external validation is ASBG03 from the student-context file. It records how often the student speaks the language of the test at home. It is not ITLANG_SA, which records the language in which the achievement test was administered. Neither ASBG03 nor the Home Resources for Learning scale ASBGHRL appears in the saved 203-feature clustering matrix; both were merged with the cluster labels only after estimation.

Post-hoc interpretation was carried out only after the cluster labels had been obtained. Clusters were characterized using weighted sizes, held-out reading plausible values, Home Resources for Learning, home-language distributions, and additional contextual indicators, together with formal tests and effect sizes. Thus, the post-hoc stage did not generate the clusters; it was used only to describe and externally assess them after estimation.

These dimensions were selected as established educational reference dimensions, not as candidate novel determinants. The interpretive question is therefore whether their joint configuration within the fitted profiles adds nuance to prior Moroccan reading research, and whether that configuration is reproducible, rather than whether the analysis rediscovers their average association with achievement.

5.4. Clustering Stage

Three groups of algorithms were applied to the cleaned dataset:

- **Classical clustering**: k-means, Agglomerative clustering, Gaussian Mixture Models.
- **Standard dimensionality-reduction baseline**: PCA+KMeans.
- **Deep clustering**: AE+k-means, Deep Embedded Clustering (DEC), Improved DEC (IDEC).

All methods received the same standardized feature matrix X as their initial input.

Sharing the same input matrix is necessary but not sufficient for a fair comparison. The classical algorithms retain the geometry of X; PCA+KMeans performs a fixed linear reduction; and the autoencoder-based methods learn a non-linear representation before assigning clusters. DEC and IDEC go further by updating that representation with a clustering loss. Their latent spaces are therefore not neutral evaluation domains but part of the fitted models themselves.

5.5. Evaluation Protocol

Cluster quality was assessed using: Silhouette[22], Calinski-Harabasz[23], Davies-Bouldin[24]. For the primary head-to-head comparison, all three indices were computed in the same standardized input space X. For deep methods, the indices were additionally computed in the corresponding learned space Z, but only as representation-specific diagnostics. t-SNE visualizations were likewise separated into a common embedding of X and method-specific embeddings of Z.

Fair cross-space comparison. We use a three-level evaluation protocol. First, the primary head-to-head comparison computes every method's indices from its labels and distances in the same standardized input space X. Second, scores computed in a learned or reduced representation are reported only as within-representation diagnostics; they describe how well a method has organized its own fitted space and are not used to rank that method against an algorithm evaluated in X or in another representation. Third, held-out educational variables and assignment stability are used to assess whether geometric separation has utility beyond the space on which it was obtained. This ordering prevents an optimized latent geometry from being treated as intrinsically superior merely because it produces larger internal-index values.

Validation beyond internal indices. The validation strategy contains three complementary components. Educational external validity is assessed with five overall-reading plausible values, the Home Resources for Learning scale, and home-language frequency, all excluded from the clustering matrix. Held-out reading results are reported in Table 8 and Figure 8, while the weighted cluster profiles and home-language distributions are presented in Table 9 and Figure 9. Formal tests and effect sizes are reported in Table 10 rather than relying on statistical significance alone. Assignment stability is assessed with pairwise adjusted Rand indices across five random initializations and summarized with the external evidence in Table 12, while robustness to preprocessing is examined across alternative missing-data treatments in Table 7. These analyses ask whether the profiles are educationally associated and reproducible, which internal geometric indices cannot establish. Because the held-out variables still come from the same Moroccan sample, we refer to this as external-variable validation, not independent sample validation.

Cross-country replication protocol. The full preprocessing and modelling pipeline was independently refitted in Morocco (MAR), France (FRA), Singapore (SGP), and South Africa (ZAF). Every country used the same 149 harmonized features, type-aware imputation rules, $K = 5$, latent dimension 16, architecture $d \to 256 \to 64 \to 16$, 100 autoencoder pretraining epochs, 2,000 DEC/IDEC refinement iterations, and primary seed 42. Imputation and standardization were fitted separately within each country. To control the quadratic cost for the 12,422-student South African sample, Silhouette was estimated from the same reproducible random sample size of at most 3,000 students in every country; Calinski-Harabasz and Davies-Bouldin used the complete national samples. The five

reading plausible values, Home Resources for Learning, and home-language frequency remained excluded from clustering and were linked afterward using the final student weight. This design tests whether the pattern of results recurs under an identical protocol, while allowing the substantive profiles and effect sizes to vary by country. The resulting cross-country clustering metrics are reported in Table 17, the corresponding held-out reading and home-resource effects in Table 18, and their comparative summary in Figure 10.

5.6. Choice of the Number of Clusters and Clustering Quality

The type-aware matrix was examined for $K \in \{3, 4, 5\}$ using the three classical procedures in the common standardized input space X. Table 1 shows that no value of K produces strong separation: the best Silhouette value is 0.116 for average-linkage agglomerative clustering at $K = 3$, and the remaining values are close to zero. We retained $K = 5$ as an exploratory resolution that permits more detailed profile description and direct comparison with the original analysis, not because the internal indices identify five natural populations. Deep models were then estimated at this fixed resolution; their latent-space scores are evaluated separately from common-space evidence below. The range $\{3, 4, 5\}$ was inherited from the original exploratory design and was not a comprehensive search over the number of profiles. We did not select K by GMM BIC or by a stability criterion over a wider range. Consequently, no claim is made that $K = 5$ is model-selected or uniquely supported; a wider prespecified K search with model-based and stability criteria remains necessary before treating the solution as a taxonomy.

Table 1. Sensitivity of classical methods to K.

Method	K	Silhouette	Calinski-Harabasz	Davies-Bouldin
k-means (X)	3	0.0452	284.83	4.0993
	4	0.0332	242.66	4.0771
	5	0.0395	211.99	4.1652
GMM full (X)	3	0.0453	283.79	4.1202
	4	0.0526	205.71	3.7904
	5	0.0523	184.58	3.7481
Agglomerative average (X)	3	0.1163	52.15	1.2354
	4	0.0949	76.43	1.9776
	5	0.0853	80.58	1.9218

5.7. Parameter Settings

The main configuration was not selected by maximizing latent-space internal indices, because doing so would favor the representation optimized by DEC/IDEC and would add post-hoc selection bias. The reported values were treated as conventional reference settings, not as a universally optimal configuration, and PCA was matched to the autoencoder latent dimension for comparability. Their influence is examined through the ablation protocol specified below and the corresponding results reported later.

5.7.1. Classical Clustering Methods

- k-means. `KMeans(n_clusters=5, n_init=20, random_state=42)` Euclidean distance and k-means++ initialization. Other parameters at scikit-learn defaults.
- Agglomerative Clustering: `AgglomerativeClustering(n_clusters=5, linkage="average")` Average linkage with Euclidean distance.
- **GMM:** `GaussianMixture(n_components=5, covariance_type="full", reg_covar=`10^{-4}`, max_iter=500, random_state=42, init_params="kmeans")` Full covariance matrices with regularization to avoid singular covariance estimates.

5.7.2. PCA+KMeans Baseline

PCA was fitted to the same standardized matrix X using 16 components, equal to the autoencoder latent dimension. K-means was then fitted to the PCA scores with $K = 5$, `n_init=20`, and seeds 0, 7, 42, 123, and 2024. Internal indices were computed both in the PCA space and, using the same labels, in X. Assignment stability was measured by pairwise ARI, and the seed-42 solution underwent the same held-out reading, home-resources, and home-language validation as DEC and IDEC.

5.7.3. Deep Clustering Methods

- Autoencoder architecture. Encoder: $d \to 256 \to 64 \to 16$. Decoder: $16 \to 64 \to 256 \to d$. ReLU activations, no dropout or batch normalization.
- AE pretraining. 100 epochs, batch size 256, Adam (lr = 10^{-3}), MSE reconstruction loss.
- AE+k-means. The pretrained encoder generates latent codes Z. k-means is applied in latent space with $k = 5$, $n_{\text{init}} = 20$.
- DEC. Cluster centers initialized by k-means in latent space. Student-t soft assignments with $\alpha = 1$. Target distribution updated at each full-data optimization step. Training for 2000 iterations, KL divergence objective, Adam with lr = 10^{-3}.
- IDEC. Same initialization as DEC but adds a reconstruction penalty: $\mathcal{L}_{\text{IDEC}} = \mathcal{L}_{\text{KL}} + \gamma \mathcal{L}_{\text{recon}}, \qquad \gamma = 1.$ Trained for 2000 full-data iterations with Adam (lr = 10^{-3}).
- t-SNE visualization. `TSNE(n_components=2, perplexity=35, random_state=42)` All other parameters at default.

6. Results and Discussion

6.1. Experimental Setup

All experiments were run on a Dell G16 7620 workstation (Core i7-12700H, 16GB RAM, RTX 3060 6GB, Windows 11 Pro 64-bit). All models were trained and evaluated on the full Moroccan PIRLS dataset. The primary autoencoder, DEC, and IDEC experiments used the RTX 3060 through CUDA, whereas the classical methods used the CPU. This mixed execution reflects the actual analysis pipeline and is not used for the computational comparison. The separate computational benchmark reported below deliberately forced every method, including the neural models, onto the CPU so that its timing ratios compare the algorithms on the same processing device rather than giving the deep methods a hardware-specific GPU advantage.

6.2. Reproducibility Details

To improve reproducibility, the revised manuscript documents the computing environment, software versions, preprocessing decisions, model configurations, hardware allocation, and random-seed management used in the analyses. The same recorded configurations were used to generate the reported clustering labels, validation statistics, sensitivity analyses, figures, and tables. The official source files must be obtained from the PIRLS 2021 International Database under its applicable access conditions; the analyses retain the original PIRLS variable names and do not replace the source data with synthetic inputs. The complete experimental code, notebooks, aggregate results, environment specification, and data-access instructions are publicly available in the accompanying GitHub repository.

Table 2. Software environment used for the primary neural analysis.

Software	Version	Software	Version
Python	3.8.20	NumPy	1.24.4
pandas	2.0.3	SciPy	1.10.1
scikit-learn	1.3.2	Matplotlib	3.7.5
PyTorch	2.4.1+cu124	pyreadstat	1.2.8

The primary reported run uses seed 42. Seed management is applied to Python's `random` module, NumPy, PyTorch, shuffled data loaders, KMeans/GMM initialization, metric subsampling, and stochastic visualizations whenever applicable. Assignment stability is evaluated independently with seeds $\{0, 7, 42, 123, 2024\}$ rather than inferred from the primary run. KMeans uses 20 initializations. The code records configurations and labels in method-specific output directories. The fresh five-seed and `IDCLASS`-exclusion reruns used a frozen CPU environment (Python 3.11.6, PyTorch 2.11.0, NumPy 2.3.3, and scikit-learn 1.8.0) with the same saved matrix, architecture, optimizer, and iteration counts. A subsequent seed-42 audit recovered the original environment reported in Table 2 and the RTX 3060. With the same saved matrix and code, that CUDA audit reproduced the archived DEC and IDEC partitions exactly (ARI $= 1.000$ for both methods in two repeated fits). Running the same original software environment on CPU instead reproduced the earlier discrepancies (ARI $= 0.456$ for DEC and 0.492 for IDEC). Thus, the discrepancy was attributable to cross-device numerical trajectories, not to a change in the preprocessing matrix, seed, initialization, architecture, optimizer, or training schedule. Exact wall-clock values and floating-point trajectories remain hardware dependent; CUDA deterministic algorithms were not enabled in the original run.

6.3. Clustering Performance Evaluation

To assess the quality and validity of the clustering results, three evaluation metrics were used: the Silhouette Coefficient, the Calinski-Harabasz Index, and the Davies-Bouldin Index.

Silhouette Coefficient. The Silhouette coefficient measures the similarity of a data point to its own cluster compared to other clusters. It can be computed based on the following formula:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where $a(i)$ denotes the average intra-cluster distance of the data point i, and $b(i)$ denotes the minimum average distance between data point i and other points in another cluster. The average Silhouette score over all data points can then be computed.

If the Silhouette score is close to 1, the observation is much closer to its assigned cluster than to the nearest alternative cluster. A value close to 0 indicates overlap or a boundary observation. A negative value indicates that the observation is, on average, closer to another cluster than to its assigned cluster.

Calinski-Harabasz Index. The Calinski-Harabasz Index, or Variance Ratio Criterion, can be used to assess the performance of the clustering results. The formula for the Calinski-Harabasz Index can be described as follows:

$$CH = \frac{\mathrm{Tr}(B_k)}{\mathrm{Tr}(W_k)} \times \frac{N - k}{k - 1}$$

where B_k and W_k denote the between-cluster dispersion and within-cluster dispersion, respectively, and N and k denote the total number of samples and the number of clusters, respectively.

The higher the Calinski-Harabasz Index, the more well-defined, compact, and separated the resulting clusters.

Davies-Bouldin Index. The Davies-Bouldin Index can be used to assess the quality of the resulting clusters. The Davies-Bouldin Index can be computed based on the following formula:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{s_i + s_j}{d_{ij}} \right)$$

where s_i and s_j denote the average intra-cluster distances of cluster i and cluster j, respectively, and d_{ij} denotes the average distance between cluster i and cluster j.

The smaller the Davies-Bouldin Index, the higher the quality of the resulting clusters.

These three indices can be used as complementary criteria to validate the quality of the clustering results:

- High Silhouette Index (> 0.5): clusters are comparatively compact and well separated.
- High Calinski-Harabasz Index: It means the clusters have high variance between clusters compared to variance within clusters.
- Low Davies-Bouldin Index: It means the clusters are compact and well-separated.

6.4. Clustering in the Original Feature Space

We start by evaluating all the methods for clustering directly in the standardized feature space X, which includes the preprocessed PIRLS variables. The reason for doing so is to ensure a fair and methodologically consistent comparison of all the algorithms by evaluating them in a common space for classical and deep methods alike.

Method	Silhouette	Calinski–Harabasz	Davies–Bouldin
KMeans	0.0395	**211.99**	4.1652
Agglomerative	**0.0853**	80.58	**1.9218**
GMM	0.0523	184.58	3.7481
AE+KMeans	0.0322	175.05	4.6704
DEC	0.0221	177.19	4.6934
IDEC	0.0245	168.69	4.7467

Table 3. Common-space clustering metrics.

Table 3 provides the fair common-space comparison: every set of labels, including those learned by DEC and IDEC, is evaluated in the same standardized input space X. All Silhouette values remain low ($0.0221 - 0.0853$), indicating extensive overlap. Agglomerative clustering has the highest Silhouette value (0.0853), but this value still indicates weak separation. DEC and IDEC attain values of 0.0221 and 0.0245, respectively, showing that their assignments are also strongly overlapping when examined through the observed variables. At this stage, the result supports only a common-space conclusion: none of the six methods produces clearly separated groups in X.

6.5. Cluster Visualization in the Standardized Input Space

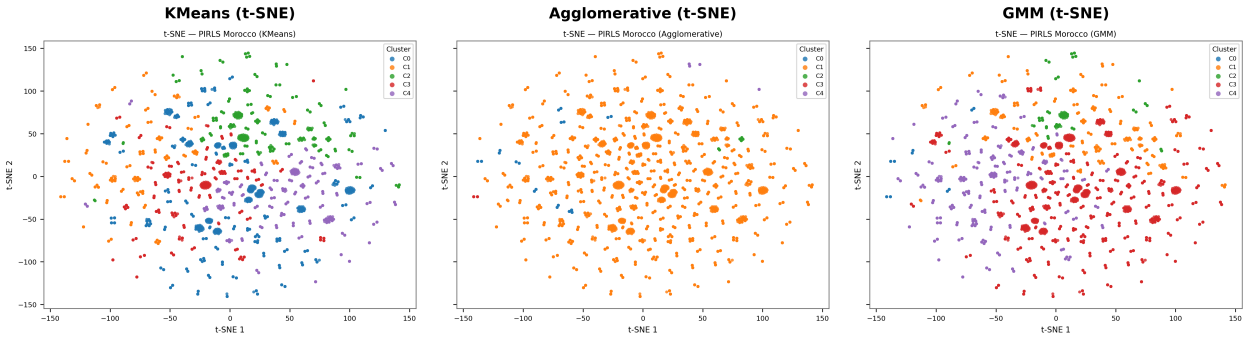


Figure 4. t-SNE visualization of classical labels in the standardized input space.

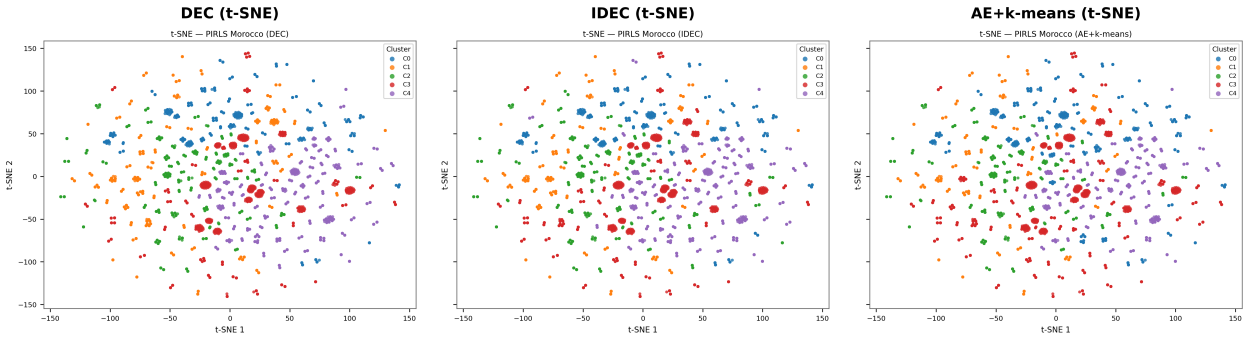


Figure 5. t-SNE visualization of deep-method labels in the standardized input space.

Figures 4 and 5 display a common t-SNE embedding of the standardized PIRLS input space X. The point coordinates are identical across all six panels; only the colors, corresponding to the cluster labels produced by each method, change. This common embedding facilitates a descriptive comparison of how the six partitions are distributed over the observed feature space. Because t-SNE is a nonlinear two-dimensional visualization that can distort global distances, it is used only for exploratory inspection and not as formal evidence of cluster validity.

The labels produced by KMeans, agglomerative clustering, and GMM remain substantially intermixed. The deep-method labels produced by AE+KMeans, DEC, and IDEC are also dispersed across overlapping regions when displayed on the same embedding of X. In particular, the visualizations do not show five clearly separated regions for the deep methods. This pattern is consistent with the low common-space Silhouette values reported in Table 3, which range from 0.0221 to 0.0853. Thus, the figures provide descriptive support for substantial overlap in the observed variables, but they do not by themselves prove the absence of latent structure.

7. Deep Clustering in the Latent Space

Deep methods transform the input matrix X into a representation Z whose geometry is influenced by the training objective. Table 4 therefore reports each primary partition in both spaces. AE+KMeans yields only modest separation in Z. DEC, by contrast, reaches a Silhouette value of 0.9994 in Z but only 0.0221 in X; IDEC reaches 0.8845 in Z and 0.0245 in X. The sharper latent geometry is thus not preserved in the observed-feature representation and is interpreted as optimization-induced rather than as evidence of ground-truth categories. All five DEC clusters remain populated, so this is not collapse into a single cluster; nevertheless, the near-perfect latent score indicates objective-induced over-separation and overly confident assignments and is therefore treated as a limitation rather than evidence of superior educational validity.

Table 4. Deep-method metrics in latent and input spaces.

Method	Space	Silhouette	Calinski-Harabasz	Davies-Bouldin
AE+KMeans	Z	0.1106	681.54	2.3597
AE+KMeans	X	0.0322	175.05	4.6704
DEC	Z	0.9994	3.49×10^9	0.0008
DEC	X	0.0221	177.19	4.6934
IDEC	Z	0.8845	134775.6	0.1635
IDEC	X	0.0245	168.69	4.7467

To distinguish the fixed primary run from stochastic variability, Table 5 summarizes the internal indices obtained in the fresh five-seed analysis (seeds 0, 7, 42, 123, and 2024). The near-perfect DEC latent Silhouette is reproducible as a property of the optimized representation, but its common-space Silhouette remains near zero. The variability of the index values should not be confused with assignment stability: as shown later in Table 11, partitions can have similar index values while assigning individual students differently.

All five seeds, including seed 42, were freshly trained on CPU with the final saved matrix, architecture, optimizer, epoch count, and refinement settings. The CPU seed-42 assignments differ from the archived CUDA partitions (ARI $= 0.456$ for DEC and 0.492 for IDEC), and their associated metrics differ slightly from Table 4. A targeted audit using the recovered original Python 3.8/PyTorch 2.4.1 CUDA environment and RTX 3060 reproduced both archived partitions exactly (ARI $= 1.000$ in two repeated fits). The same original software run on CPU yielded ARI $= 0.456$ and 0.492, isolating execution device as the source of the partition discrepancy. Tables 5 and 11 deliberately retain the complete CPU five-seed sensitivity analysis, whereas the primary tables retain the exactly reproduced CUDA seed-42 partitions used by the profile analyses. Small run-to-run differences remain in the continuous CUDA latent coordinates because deterministic CUDA algorithms were not enabled, although the hard seed-42 partitions were reproduced exactly.

Table 5. Deep-method internal indices across five random seeds (mean $\pm$ SD).

Method	Space	Silhouette	Calinski–Harabasz	Davies–Bouldin
DEC	Z	0.9996 ± 0.0001	$(5.85 \pm 1.56) \times 10^9$	0.0006 ± 0.0001
DEC	X	0.0238 ± 0.0020	181.23 ± 17.23	4.4722 ± 0.3515
IDEC	Z	0.8858 ± 0.0029	124457.29 ± 9885.33	0.1606 ± 0.0036
IDEC	X	0.0253 ± 0.0030	181.63 ± 18.59	4.4597 ± 0.3402

7.1. Visualization of the Latent Spaces

Figure 6 displays a separate t-SNE projection for each 16-dimensional latent representation. Unlike Figures 4 and 5, which use one common embedding of the standardized input matrix X, each panel here is fitted directly to the method-specific representation Z. The panels therefore illustrate the geometry created by each encoder and objective, but distances and apparent compactness must not be compared quantitatively across panels. AE+KMeans retains substantial overlap, whereas DEC and IDEC create much sharper latent regions, consistently with their latent-space internal indices. This visual separation is treated as a diagnostic of representation learning rather than evidence of naturally discrete student populations.

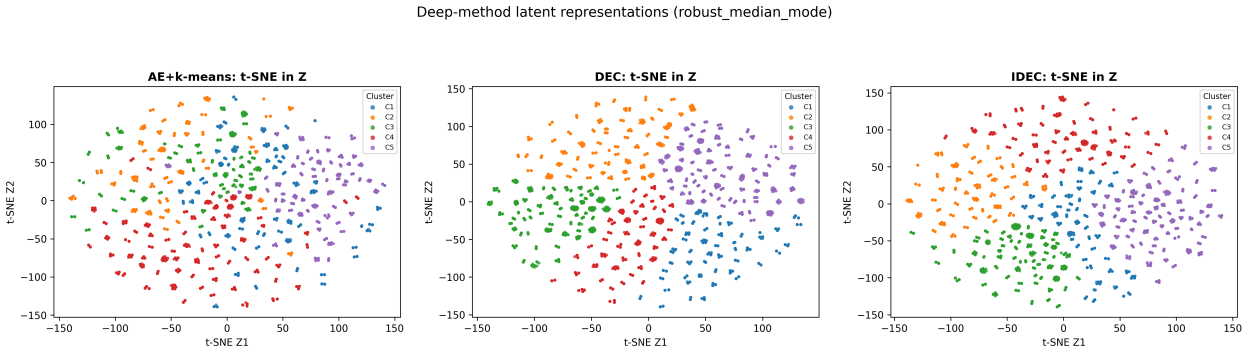


Figure 6. Separate t-SNE projections of the learned latent spaces Z for AE+KMeans, DEC, and IDEC under the primary type-aware preprocessing. Colors denote the five method-specific clusters.

The apparent compactness of DEC and IDEC must not be confused with evidence that the same groups naturally exist in the educational measurements. Their encoders are not neutral projections: DEC's sharpened target distribution and KL-divergence objective jointly move representations and reinforce confident assignments, while IDEC only partly constrains this deformation through reconstruction. Consequently, a continuous, overlapping input manifold can be mapped to visually and numerically separated latent regions even in the absence of ground-truth classes. Neither a near-perfect latent Silhouette nor a t-SNE plot validates ontological student types. In this study, weak common-space separation, low cross-seed assignment agreement, sensitivity to preprocessing and architecture, and modest external effect sizes all argue against such an interpretation. The latent partitions are

therefore treated as model-dependent exploratory summaries, not naturally occurring populations or a basis for individual educational decisions.

7.2. Fairness and Implications of Cross-Space Evaluation

Internal validity indices are functions of both the cluster labels and the distances in the space where they are calculated. They are not invariant to a learned non-linear transformation, changes in scale, or changes in dimensionality. A numerical comparison such as "DEC in Z versus k-means in X" therefore changes two elements at once: the partitioning procedure and the feature geometry. It cannot isolate an advantage of the clustering algorithm itself.

This issue gives deep clustering an inherent advantage in its own representation. The DEC objective explicitly moves embeddings toward centroids and sharpens confident assignments by minimizing KL divergence to an auxiliary target distribution. IDEC uses the same clustering pressure while constraining it with reconstruction. Because the distances subsequently used by Silhouette, Calinski-Harabasz, and Davies-Bouldin have already been shaped by these objectives, strong latent scores are partly an expected in-sample consequence of representation learning. In contrast, classical methods are evaluated on a fixed geometry that they cannot reshape. Even AE+KMeans and PCA+KMeans should be interpreted within their own reduced spaces with caution, because compression can remove variance that is unfavorable to separation.

Accordingly, latent-space values answer the limited question of whether a fitted model constructed a compact representation; they do not answer whether it discovered a better partition of the observed students than a classical method. The latter question is addressed here by the common-X comparison, assignment stability, and held-out educational validation. Under this fairer evidence hierarchy, the results do not show superior clustering quality for DEC or IDEC: their labels have lower Silhouette and Calinski-Harabasz and higher Davies-Bouldin values than several simpler alternatives in X. We retain the latent results for transparency and to document model behavior, but we do not use their magnitude as evidence of deep-method superiority.

7.3. Comparison with PCA+KMeans

Table 6 addresses whether the deep clustering objective adds value beyond standard dimensionality reduction. The 16 PCA components explain 44.56% of the variance. For the primary seed, PCA+KMeans attains a Silhouette of 0.0969 in its reduced space, with Calinski-Harabasz 561.51 and Davies-Bouldin 2.3813. When the same labels are evaluated in the common input space X, the scores are 0.0413, 223.67, and 3.8040, respectively. Its input-space Silhouette remains low and therefore does not establish five natural groups. Nevertheless, PCA+KMeans outperforms AE+KMeans, DEC, and IDEC on all three common-space indices, and it has the highest common-space Calinski-Harabasz value among the methods considered. Average-linkage agglomerative clustering still has the highest Silhouette and lowest Davies-Bouldin values, showing that no single method dominates every criterion.

Table 6. Common-space comparison including PCA+KMeans.

Method	Silhouette$_X$	Calinski-Harabasz$_X$	Davies-Bouldin$_X$
KMeans	0.0395	211.99	4.1652
Agglomerative	**0.0853**	80.58	**1.9218**
GMM	0.0523	184.58	3.7481
PCA+KMeans	0.0413	**223.67**	3.8040
AE+KMeans	0.0322	175.05	4.6704
DEC	0.0221	177.19	4.6934
IDEC	0.0245	168.69	4.7467

PCA+KMeans is also considerably more stable across the five random initializations: its mean pairwise ARI is 0.702 (range 0.579-0.840), compared with 0.221 for DEC and 0.222 for IDEC in the fresh five-seed rerun. Its seed-42 partition differs substantially from the archived primary deep partitions (ARI 0.173 with AE+KMeans, 0.162 with DEC, and 0.161 with IDEC), so the methods are not simply recovering the same groups under different representations. Held-out validation shows statistically detectable profile differences for reading across all five

plausible values (maximum $p = 3.69 \times 10^{-16}$), with a modest weighted effect size ($\eta^2 = 0.025$). Home resources produce a larger contrast ($\eta^2 = 0.067$), whereas home language remains weakly associated with membership (Cramer's $V = 0.102$). Thus, the simpler PCA baseline is more stable and at least as competitive geometrically, while the deep solutions show a somewhat larger reading contrast ($\eta^2 \simeq 0.045$-0.049). These mixed results do not support a general claim that the clustering-specific deep objectives are necessary.

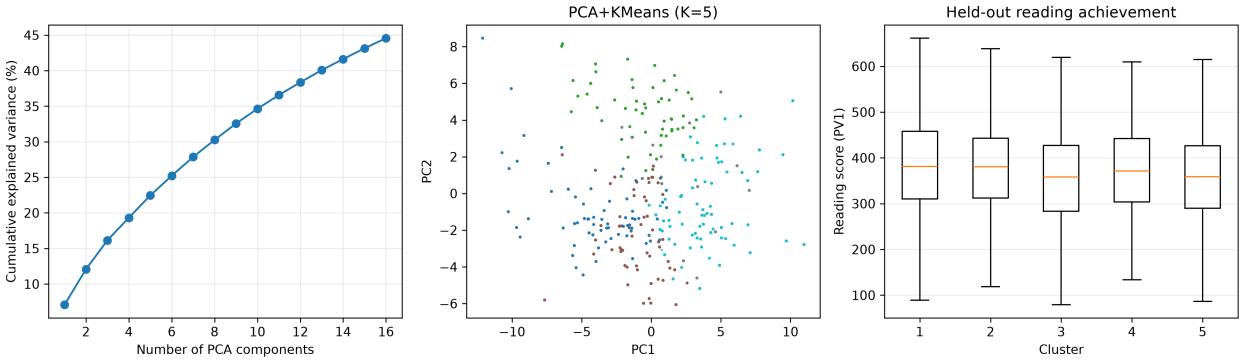


Figure 7. PCA+KMeans baseline validation. Left: cumulative variance explained by the 16 retained components. Center: seed-42 labels projected on the first two principal components. Right: descriptive distribution of held-out reading achievement (PV1) by cluster; formal reading comparisons use all five plausible values.

Figure 7 complements the numerical comparison. The left panel shows that the 16 retained components explain 44.56% of the cumulative variance. The center panel projects the five-cluster solution onto the first two principal components and shows substantial overlap; because clustering used all 16 components, this two-dimensional view is descriptive rather than a complete assessment of separation. The right panel likewise shows overlapping PV1 reading distributions across clusters, although formal tests using all five plausible values detect modest profile differences. Thus, PCA+KMeans retains some educational differentiation without providing evidence of five sharply separated natural groups.

7.4. Sensitivity to Missing-data Treatment

Table 7 compares student-level assignments obtained from the former median-only specification with those from the type-aware primary analysis, and then compares the primary assignments with the first stochastic iterative-imputation dataset. The results show substantial preprocessing sensitivity. The ARI between the previously saved and primary type-aware partitions ranges from 0.030 for GMM to 0.282 for *k*-means; for DEC and IDEC it is 0.208 and 0.180, respectively. Comparison of the type-aware and iterative-imputation partitions remains limited, with ARI values between 0.191 and 0.423 across the six methods. These values mean that similar aggregate indices do not imply that the same students remain grouped together.

Table 7. Assignment sensitivity to missing-data treatment.

Method	**Former vs. primary**	**Primary vs. iterative-1**
k-means	0.282	0.219
Agglomerative	0.224	0.191
GMM	0.030	0.253
AE+KMeans	0.135	0.356
DEC	0.208	0.423
IDEC	0.180	0.379

Across all five stochastic imputations, pairwise *k*-means agreement ranges from 0.254 to 0.825 (mean ARI $= 0.486$), indicating that imputation uncertainty can materially alter membership. Nevertheless, the central cross-space finding remains unchanged. With type-aware preprocessing, DEC attains a latent-space Silhouette of 0.9994

but only 0.0221 when the same labels are evaluated in the standardized input space; the corresponding values under the first iterative imputation are 0.9941 and 0.0320. For IDEC, the values are 0.8845 versus 0.0245 under type-aware preprocessing and 0.8803 versus 0.0309 under iterative imputation. Therefore, alternative missing-data treatments do not support a claim of natural separation in the observed feature space. Instead, they add a further limitation: the identities of individual cluster members are preprocessing-dependent.

7.5. External Educational Validation

To assess utility rather than relying only on internal geometry, we linked the cluster assignments to the five PIRLS overall-reading plausible values (ASRREA01-ASRREA05), which were held out from the clustering matrix. Estimates used the final student weight. For the primary seed (42), uncertainty was estimated by resampling schools and then combining the five plausible-value estimates; Table 8 reports the resulting weighted means and 95% intervals. Cluster numbers are descriptive labels and are not assumed to match across methods or seeds.

Table 8. Held-out reading achievement by cluster.

	DEC		IDEC	
Cluster	n	Weighted mean (95% interval)	n	Weighted mean (95% interval)
1	1301	365.35 (348.07, 382.62)	1209	364.10 (345.41, 382.78)
2	1565	369.96 (355.45, 384.47)	1290	368.00 (350.79, 385.22)
3	1216	342.09 (327.10, 357.08)	1324	345.29 (330.94, 359.63)
4	1312	355.94 (343.88, 368.00)	1513	356.42 (345.56, 367.29)
5	1623	402.24 (379.64, 424.85)	1681	401.12 (378.95, 423.30)

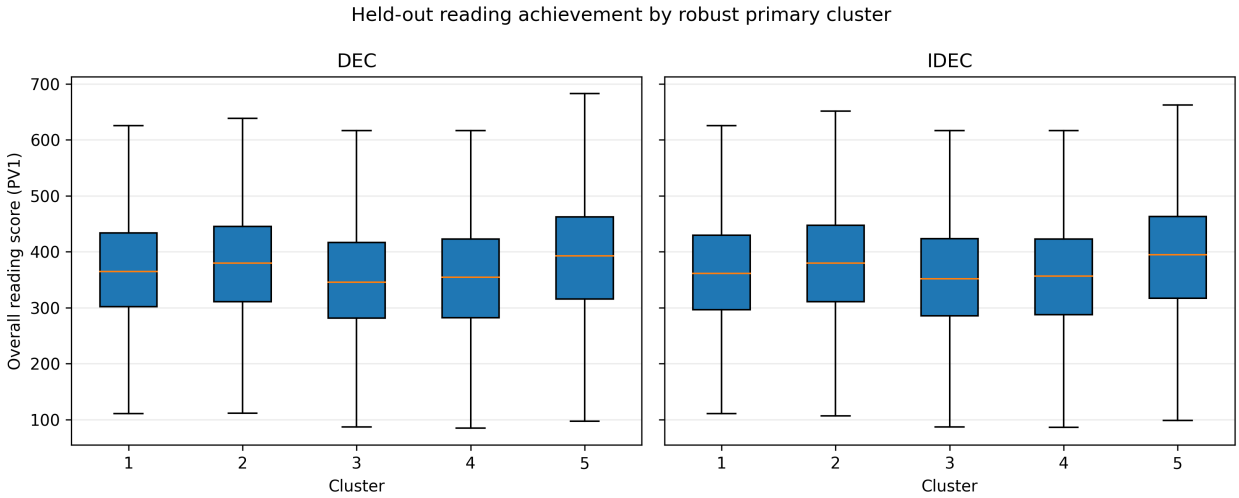


Figure 8. Distribution of the held-out overall reading score (PV1) by cluster. The box plots are descriptive; formal comparisons pool evidence across all five plausible values.

Held-out reading achievement provides evidence of external validity, although the association is moderate rather than decisive. For both methods, Cluster 5 had the highest weighted reading mean (402.24 for DEC and 401.12 for IDEC), whereas Cluster 3 had the lowest mean (342.09 and 345.29, respectively), corresponding to differences of 60.15 and 55.83 PIRLS points. Kruskal-Wallis tests [25] conducted separately for the five plausible values consistently rejected equality of the score distributions across clusters: the statistics ranged from 127.66 to 151.44 for DEC and from 112.94 to 144.09 for IDEC, with all p-values no greater than 1.72×10^{-23}. Nevertheless, the pooled effect sizes were only 0.049 for DEC and 0.045 for IDEC, and Figure 8 shows substantial overlap between the distributions. These results indicate that cluster membership is associated with reading achievement, but that

the profiles describe partially differentiated positions along an achievement gradient rather than sharply separated performance groups.

7.6. Quantitative Cluster Profiles

To replace the earlier qualitative interpretation, we use the Home Resources for Learning scale (ASBGHRL) and frequency of speaking the test language at home (ASBG03), neither of which was included in the clustering matrix. These variables characterize, rather than define, the fitted profiles. Percentages and means use the PIRLS final student weight. Because cluster labels are model-specific, DEC and IDEC profiles are reported separately and cluster numbers are not interpreted as ordered levels. ASBGHRL measures home learning resources and is not treated as a comprehensive measure of socioeconomic status.

Table 9. Profiles based on held-out educational variables.

Method	Cluster	Weighted share (%)	Reading mean	HRL mean	Always/almost (%)	Sometimes (%)	Never (%)
DEC	1	10.83	365.35	6.86	38.71	22.73	38.57
DEC	2	18.97	369.96	7.63	51.82	25.22	22.95
DEC	3	22.40	342.09	6.70	29.95	21.50	48.55
DEC	4	13.52	355.94	6.95	30.64	33.11	36.25
DEC	5	34.28	402.24	7.92	46.65	29.14	24.21
IDEC	1	10.61	364.10	6.89	41.66	22.20	36.14
IDEC	2	16.71	368.00	7.62	48.87	27.57	23.55
IDEC	3	22.32	345.29	6.75	33.46	21.31	45.24
IDEC	4	15.23	356.42	7.01	32.32	30.33	37.35
IDEC	5	35.13	401.12	7.87	45.05	29.02	25.93

For the primary DEC solution, Cluster 3 contains 22.40% of the weighted population and combines the lowest reading mean (342.09) and HRL mean (6.70) with the largest percentage reporting never speaking the test language at home (48.55%). Cluster 2 represents 18.97% and has the largest frequent-use percentage (51.82% always or almost always). Cluster 5 is the largest (34.28%) and has the highest reading (402.24) and HRL (7.92) means, while 24.21% report never speaking the test language at home. Cluster 1 (10.83%) is intermediate in reading but combines relatively limited HRL with a mixed home-language distribution, whereas Cluster 4 (13.52%) is a lower-intermediate profile with the largest percentage reporting that they sometimes speak the test language at home (33.11%).

The IDEC profiles reproduce the broad descriptive pattern without implying identical student membership. Cluster 1 (10.61%) is an intermediate profile with relatively low HRL and a mixed language distribution. Cluster 2 (16.71%) combines relatively high HRL with the largest frequent-use percentage (48.87%). Cluster 3 (22.32%) has the lowest reading (345.29) and HRL (6.75) means and the largest never-speaking percentage (45.24%). Cluster 4 (15.23%) is a lower-intermediate profile with comparatively frequent occasional use of the test language (30.33%). Cluster 5 is the largest (35.13%) and combines the highest reading (401.12) and HRL (7.87) means. These are descriptive associations rather than causal effects or naturally bounded student types.

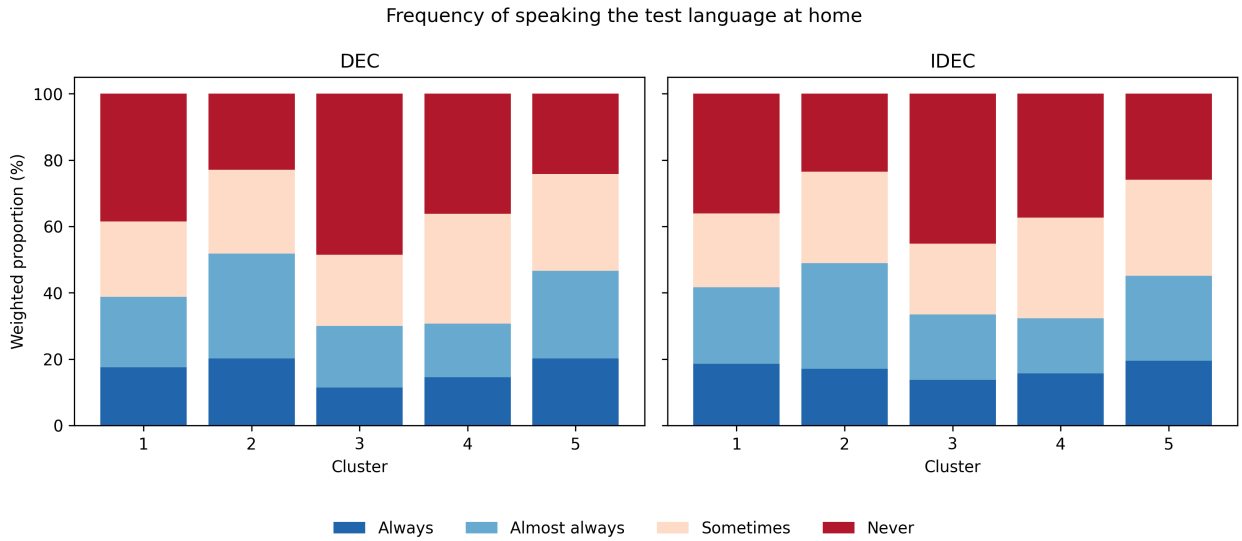


Figure 9. Weighted distribution of frequency of speaking the test language at home across the five clusters.

Table 10. Tests and effect sizes for held-out variables.

Outcome	Method	Test statistic	Exploratory p-value	Effect size
Reading (PV1-PV5)	DEC	$H = 127.66$-151.44	$p_{\max} = 1.23 \times 10^{-26}$	weighted $\eta^2 = 0.049$
Reading (PV1-PV5)	IDEC	$H = 112.94$-144.09	$p_{\max} = 1.72 \times 10^{-23}$	weighted $\eta^2 = 0.045$
Home Resources for Learning	DEC	$H = 170.37$	8.70×10^{-36}	weighted $\eta^2 = 0.078$
Home Resources for Learning	IDEC	$H = 143.22$	5.78×10^{-30}	weighted $\eta^2 = 0.067$
Home-language frequency	DEC	$\chi^2 = 274.41$	1.09×10^{-51}	Cramer's $V = 0.121$
Home-language frequency	IDEC	$\chi^2 = 187.20$	1.42×10^{-33}	Cramer's $V = 0.100$

Table 10 makes the formal comparisons explicit. The Kruskal-Wallis and chi-square tests are conventional unweighted exploratory tests; their p-values are not design-based PIRLS population inferences. We therefore give greater interpretive weight to the weighted descriptive profiles and weighted η^2 estimates. Home resources show a consistent contextual gradient (weighted $\eta^2 = 0.078$ for DEC and 0.067 for IDEC). Home-language distributions also differ statistically, but the conventional Cramer's V values [26] are small (0.121 and 0.100). Because η^2 and Cramer's V quantify different types of association, their numerical values are not compared as if they were the same scale. Figure 9 displays the weighted language distributions explicitly. Taken together, the results support limited educational differentiation, not a claim that home language or home resources causally determine cluster membership.

7.7. Assignment Stability

We also compared individual assignments across the five random seeds using the adjusted Rand index (ARI), which is invariant to permutations of cluster labels. Mean pairwise ARI is only 0.221 for DEC and 0.222 for IDEC (ranges 0.143-0.330 and 0.113-0.336, respectively), as detailed in Table 11. Thus, apparently stable latent-space internal scores do not imply that the same students are consistently grouped together. Taken jointly, the external results support a limited use of the clusters as exploratory summaries associated with achievement and home resources, but not as a unique, stable classification or a direct basis for individual or policy decisions. As noted with Table 5, these values come from the fresh complete CPU rerun and therefore quantify cross-seed stability under one common execution device. The targeted rerun in the recovered original CUDA environment exactly reproduced the archived primary seed-42 partitions.

Table 11. Assignment stability across five random seeds.

Method	Seeds	Pairwise comparisons	Mean ARI	SD	Range
DEC	$\{0, 7, 42, 123, 2024\}$	10	0.221	0.064	0.143-0.330
IDEC	$\{0, 7, 42, 123, 2024\}$	10	0.222	0.073	0.113-0.336

7.8. Integrated Evidence Beyond Internal Metrics

Table 12 brings the non-geometric evidence together. For DEC and IDEC, held-out reading achievement and home resources differ across profiles, but the effect sizes are modest and the assignments are unstable across initializations. PCA+KMeans is substantially more stable, while its associations with held-out variables are of comparable order. This pattern is important: extremely high latent-space indices for DEC do not translate into uniquely stronger educational validity or assignment reproducibility. The profiles therefore have some descriptive educational relevance, but the evidence does not establish practical superiority of the deep partitions.

Table 12. External validity and assignment-stability summary.

Method	Reading η^2	Home-resources η^2	Home-language V	Seed ARI
PCA+KMeans	0.025	0.067	0.102	0.702
DEC	0.049	0.078	0.121	0.221
IDEC	0.045	0.067	0.100	0.222

These analyses address external-variable validity and initialization stability, but they do not constitute independent replication: all held-out outcomes belong to the same PIRLS Moroccan sample. We also did not conduct expert adjudication or a leakage-free out-of-sample predictive comparison with and without cluster membership. Consequently, we do not claim that the clusters improve prediction or are ready for intervention and policy use. Independent-country replication, expert assessment, and nested predictive evaluation are treated as necessary subsequent tests rather than inferred from the internal indices or the present post-hoc associations.

7.9. Educational Interpretation: Confirmation and Refinement

The association of family resources, language background, and contextual conditions with reading achievement is already well established in PIRLS research and has also been reported for Moroccan fourth-grade students using supervised machine learning [9]. We therefore do not present the presence of these variables in the profiles as a new substantive discovery. The contribution of the clustering analysis is narrower: it tests whether familiar correlates combine into reproducible multivariate configurations and whether deep representation learning reveals a clearer organization than simpler alternatives.

Table 13. Established knowledge and profile-based refinement.

Established relationship	Evidence in the fitted profiles	Refinement and limitation
Greater home resources are generally associated with stronger reading performance.	Home resources provide the largest measured contextual contrast ($\eta^2 = 0.078$ for DEC and 0.067 for IDEC); the highest-reading DEC profile also has the highest HRL mean.	The analysis quantifies how this gradient appears inside multivariate profiles, but does not establish a new relationship or a causal effect.
Home-language exposure is related to educational experience and achievement.	Language distributions differ statistically, but Cramer's V is only 0.121 for DEC and 0.100 for IDEC. The DEC profile with the largest frequent-use percentage is not the profile with the highest reading mean.	Language exposure is not supported as the dominant organizing axis; its role is conditional on a broader configuration and weaker than suggested by significance alone.
Educational outcomes reflect multiple home and school contexts.	Reading and resource levels do not produce a single monotonic ordering across all five profiles, and substantial within-profile reading overlap remains.	The profiles summarize joint configurations rather than isolated predictors, but the available evidence cannot attribute differences to a specific contextual mechanism.
Data-driven segmentation may reveal learner heterogeneity.	DEC and IDEC have low cross-seed ARI (0.221 and 0.222), while PCA+KMeans is more stable and comparably informative on the held-out variables.	The most important refinement is negative: established gradients do not crystallize into a unique taxonomy, and deep clustering does not yield demonstrably novel learner types.

Two profile contrasts illustrate this qualified refinement. In the DEC solution, Cluster 2 has the largest weighted proportion of students who always or almost always speak the test language at home (51.82%), but its reading mean is 369.96. Cluster 5 has a lower frequent-use proportion (46.65%), yet has the highest reading mean (402.24) and the highest HRL mean (7.92). Moreover, Clusters 1 and 4 have similar HRL means (6.86 and 6.95), but different reading means (365.35 and 355.94) and frequent-language proportions (38.71% and 30.64%). These descriptive contrasts illustrate that the fitted profiles do not follow a simple ordering by home-language frequency or learning resources. They may motivate further hypotheses, but do not establish causal mechanisms or previously unknown educational determinants.

Accordingly, the substantive extension is not the discovery of previously unknown factors. It is the finding that known educational gradients are represented as overlapping and model-dependent configurations, with home resources more discriminating than home language and with no stable deep-clustering taxonomy. This result refines the scope of earlier variable-importance findings by distinguishing population-level correlates from robust person-level profiles.

7.10. Sensitivity of Classical Baselines

Table 14. Sensitivity of classical baselines to configuration.

Method	Configuration	Silhouette	Calinski-Harabasz	Davies-Bouldin
Agglomerative	average	0.0853	80.58	1.9218
Agglomerative	complete	0.0596	121.94	3.2526
Agglomerative	ward	0.0392	180.46	4.0327
GMM	diag	0.0342	162.51	4.3611
GMM	full	0.0523	184.58	3.7481
GMM	spherical	0.0513	186.37	3.6522
GMM	tied	0.0523	184.48	3.7489

Table 14 shows that the assessment of the classical baselines depends on their configuration and on the selected validity index. For agglomerative clustering, average linkage gives the best Silhouette (0.0853) and Davies-Bouldin (1.9218) values, indicating the most compact and separated solution according to these two criteria. Ward linkage instead produces the highest Calinski-Harabasz value (180.46), but its lower Silhouette and higher Davies-Bouldin value indicate substantially greater overlap. For GMM, full and tied covariance structures give the highest Silhouette value (0.0523), whereas the spherical structure gives the lowest Davies-Bouldin value (3.6522). No configuration therefore dominates every index. These results justify reporting the configuration choices explicitly and reinforce the broader conclusion that all classical solutions remain only weakly separated in the original feature space.

7.11. Hyperparameter Ablation

We evaluated 27 configurations through a one-factor-at-a-time screening ablation on a fixed random sample of 3,000 Moroccan students. The analysis varied latent dimension $\{8, 16, 32\}$, encoder architecture (shallow, base, deep), IDEC reconstruction weight $\gamma \in \{0.1, 1, 10\}$, refinement duration $\{100, 250, 500\}$, and learning rate $\{10^{-4}, 10^{-3}, 3 \times 10^{-3}\}$. Screening used 30 autoencoder epochs and at most 500 refinement iterations, whereas the main models were trained on all 7,017 students for 100 and 2,000 iterations. Every comparison used seed 42, $K = 5$, and the same sampled students. Table 15 reports the resulting common-space Silhouette ranges and the minimum ARI relative to the base screening configuration.

Table 15. Hyperparameter-ablation results.

Factor (settings)	Method	Silhouette$_X$ range	Minimum ARI	Main setting
Latent dimension (8, 16, 32)	DEC	0.0096-0.0374	0.125	16
	IDEC	0.0185-0.0344	0.142	16
Architecture (shallow, base, deep)	DEC	0.0259-0.0382	0.177	base
	IDEC	0.0300-0.0343	0.164	base
Reconstruction weight (0.1, 1, 10)	IDEC	0.0268-0.0300	0.866	1
Iterations (100, 250, 500)	DEC	0.0259-0.0259	1.000	500†
	IDEC	0.0296-0.0300	0.979	500†
Learning rate ($10^{-4}, 10^{-3}, 3 \times 10^{-3}$)	DEC	0.0250-0.0289	0.937	10^{-3}
	IDEC	0.0300-0.0303	0.955	10^{-3}

†Screening maximum; the full-data models use 2,000 refinement iterations.

The ablation qualifies the role of the design choices. Latent dimension and architecture produce the largest membership changes: relative to the 16-dimensional base solution, ARI falls to 0.125 for DEC and 0.142 for IDEC across dimensions, and to 0.177 and 0.164 across architectures. Consequently, a single partition should not be treated as architecture-invariant. The 16-dimensional base network was retained as a transparent middle-capacity reference and for dimensional parity with PCA, not because it maximized Silhouette$_X$ in the screening experiment.

By contrast, the optimization settings have limited influence on the main common-space conclusion. Changing γ from 0.1 to 10 changes IDEC's latent Silhouette from 0.942 to 0.162, but Silhouette$_X$ remains within 0.0268-0.0300. Increasing the screening duration from 100 to 500 iterations sharply increases DEC's latent Silhouette from 0.954 to 0.997 while leaving its labels and Silhouette$_X$ unchanged (ARI 1.000; Silhouette$_X = 0.0259$). Similarly, learning-rate changes alter latent compactness more than common-space validity. Thus, the selected 2,000-iteration budget ensures optimization but does not explain away the cross-space discrepancy; if anything, continued optimization mainly sharpens Z. The ablation therefore supports the paper's cautious interpretation while demonstrating that profile membership remains sensitive to representational capacity.

7.12. Computational Efficiency and Scalability

We complemented the statistical comparison with an isolated-process benchmark on the final robust Moroccan matrix ($n = 7{,}017$, $d = 203$). Each method used $K = 5$ and the same settings as the main experiment. Neural models used 100 autoencoder epochs; DEC and IDEC additionally used 2,000 full-batch refinement iterations.

Every run was performed on CPU in a fresh Python process. Wall-clock time was measured around model fitting, and working-set memory was sampled every 10 ms. “Peak memory” includes the Python runtime and loaded libraries, whereas “incremental memory” is the increase over that method’s post-import, post-data-load baseline. To examine empirical scaling, the identical protocol was run on fixed random subsamples of 1,000 and 3,000 students. The values in Table 16 are controlled single-run measurements and should be interpreted as machine-specific comparative evidence, not as universal runtime constants.

Table 16. CPU time and memory benchmark.

Method	Time 1,000 (s)	Time 3,000 (s)	Time 7,017 (s)	Time ratio	Peak memory (MB)	Incremental memory (MB)
KMeans	0.73	1.27	1.44	1.96	323.0	21.7
Agglomerative	0.05	0.32	5.23	98.31	680.0	378.6
GMM	0.57	0.70	1.44	2.52	373.6	71.3
PCA+KMeans	0.75	0.93	0.95	1.26	317.2	15.5
AE+KMeans	5.48	11.18	18.99	3.47	420.3	118.6
DEC	20.07	36.21	75.16	3.75	537.2	235.4
IDEC	25.49	52.80	102.19	4.01	539.1	237.5

The theoretical costs explain the observed trade-offs. With I optimization iterations, KMeans requires approximately $O(nKdI)$ operations and $O(nd + Kd)$ storage. Unconstrained agglomerative clustering requires pairwise information and therefore has approximately quadratic time and memory growth in n. A full-covariance GMM costs approximately $O(InKd^2)$ and stores $O(nd + Kd^2)$ values. Full-SVD PCA costs $O(\min(nd^2, n^2d))$, after which KMeans operates in the 16-dimensional representation. For the neural methods, an autoencoder epoch costs $O(nW)$, where W represents the layer-wise multiply-accumulate cost. DEC then adds encoder and K-centroid assignment operations over 2,000 iterations, while IDEC also evaluates the decoder and reconstruction loss during refinement.

On the complete dataset, PCA+KMeans, KMeans, and GMM finish in approximately one second, whereas AE+KMeans requires 18.99 s, DEC 75.16 s, and IDEC 102.19 s. Relative to KMeans, DEC and IDEC are about 52 and 71 times slower, respectively. The phase measurements show why: DEC spends 53.92 s in refinement after 18.28 s of pretraining, and IDEC spends 80.00 s in refinement after 19.73 s of pretraining. Deep models also use more incremental memory than KMeans (235-238 MB versus 22 MB), although agglomerative clustering has the largest peak and incremental requirements (680 and 379 MB) because of its pairwise structure. Its empirical time increase is also the steepest when moving from 1,000 to 7,017 students. Thus, the learned representations do not provide a computational advantage here: their possible exploratory value must be weighed against substantially longer training, while agglomerative clustering poses the clearest memory-scalability limitation. Mini-batch training, early stopping, approximate nearest-neighbor structures, and GPU execution could change absolute costs, but they were not used in this like-for-like CPU benchmark.

7.13. Cross-Country Robustness and Generalizability

For clarity, the Moroccan results in Tables 17 and 18 come from an independent refit on the 149-feature harmonized matrix and therefore are not expected to reproduce the 203-feature primary Moroccan results reported in Tables 6 and 12.

The harmonized replication evaluates whether the central Moroccan finding persists in three additional education systems. Table 17 shows a consistent cross-space divergence. DEC reaches latent Silhouette values between 0.9991 and 0.9997 in all four countries, while its common-space values range only from 0.0093 to 0.0571. IDEC shows the same pattern, with latent values of 0.8676-0.9061 and common-space values of 0.0095-0.0545. Thus, the conclusion that optimized latent geometry should not be interpreted as natural separation is not unique to Morocco.

Table 17. Cross-country clustering comparison.

Country	KMeans$_X$	PCA+KMeans$_X$	AE+KMeans$_X$	DEC$_Z$	DEC$_X$	IDEC$_Z$	IDEC$_X$
Morocco	0.0561	0.0514	0.0266	0.9995	0.0220	0.8840	0.0208
France	0.0527	0.0536	0.0461	0.9997	0.0485	0.9009	0.0459
Singapore	0.0700	0.0679	0.0544	0.9991	0.0571	0.9061	0.0545
South Africa	0.0403	0.0396	0.0246	0.9992	0.0093	0.8676	0.0095
Mean	0.0548	0.0531	0.0379	0.9994	0.0342	0.8897	0.0327

In the common space, no country exhibits strong separation. KMeans has the highest Silhouette in Morocco, Singapore, and South Africa, while PCA+KMeans is marginally higher in France. Averaged across countries, KMeans (0.0548) and PCA+KMeans (0.0531) exceed AE+KMeans (0.0379), DEC (0.0342), and IDEC (0.0327). The international evidence therefore reinforces the balanced interpretation: deep objectives reliably manufacture compact latent geometry, but they do not yield stronger separation of the observed harmonized data.

Table 18. Cross-country external-validation effect sizes.

	Reading achievement			Home resources		
Country	PCA+KMeans	DEC	IDEC	PCA+KMeans	DEC	IDEC
Morocco	0.0840	0.0832	0.0817	0.1511	0.1326	0.1193
France	0.0704	0.0560	0.0567	0.0913	0.0699	0.0721
Singapore	0.0228	0.0442	0.0469	0.0245	0.0521	0.0489
South Africa	0.1673	0.0888	0.0859	0.1853	0.1133	0.1138
Mean	0.0861	0.0681	0.0678	0.1131	0.0920	0.0885

Table 18 demonstrates substantive heterogeneity rather than a universal profile structure. Reading effects are largest in South Africa and smallest in Singapore, and the relative ordering of methods changes across countries. DEC and IDEC exceed PCA+KMeans for the Singaporean reading and home-resource effects, whereas PCA+KMeans has markedly larger effects in South Africa and France. Across all countries, mean reading and home-resource effects are larger for PCA+KMeans than for DEC or IDEC. Home-language associations remain small for every method on average (Cramer's $V \simeq 0.07$-0.08). Consequently, the framework is computationally reproducible and the cross-space warning is generalizable, but cluster composition and educational utility remain country-specific.

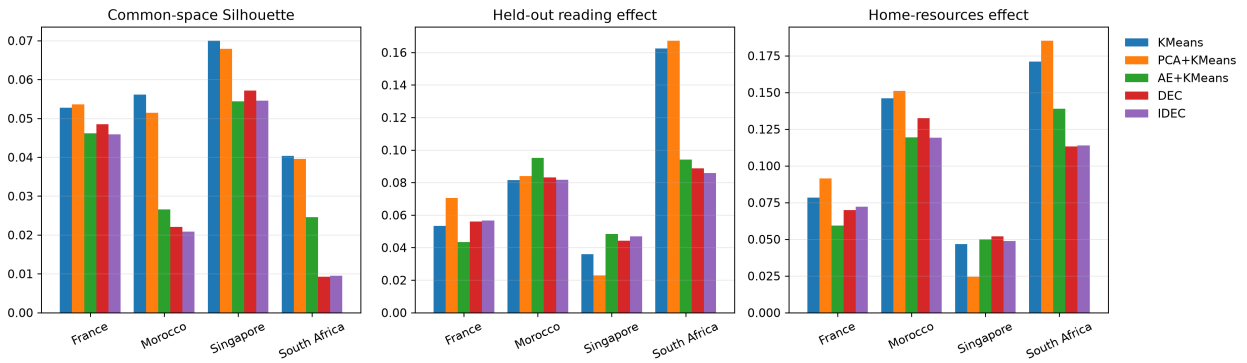


Figure 10. Cross-country comparison of common-space Silhouette, held-out reading effect, and Home Resources for Learning effect. Reading effects average the five plausible-value-specific weighted η^2 estimates.

This experiment establishes within-country replication under a harmonized protocol; it does not establish a single international learner taxonomy. Models were refitted separately rather than transferred unchanged to a held-out

country, and cluster labels are not aligned across systems. Future work should therefore add leave-one-country-out representation transfer and formal profile-centroid alignment. The present evidence is appropriately interpreted as robustness of the methodological diagnosis, not invariance of the learner profiles themselves.

7.14. Sensitivity to Excluding the Identifier-like Class Variable

The final feature audit revealed that `STU_IDCLASS`, a classroom identifier-like field, had been retained in the 203-feature primary matrix. We therefore repeated the seed-42 analysis after deleting this column, leaving 202 features. All methods were refitted; the neural methods were retrained from initialization with the same architecture, 100 pretraining epochs, and 2,000 refinement iterations. Table 19 compares the resulting assignments with same-code full-matrix fits using the adjusted Rand index and reports validation results for the 202-feature analysis.

Table 19. Sensitivity to excluding `STU_IDCLASS`; assignment ARI compares same-code CPU fits.

Method	Assignment ARI	Silhouette$_X$	Reading η^2	HRL η^2	Home-language V
KMeans	0.450	0.0353	0.0345	0.0862	0.0885
Agglomerative	0.839	0.0867	0.0533	0.1049	0.0642
GMM full	0.196	0.0426	0.0653	0.0966	0.0889
PCA+KMeans	0.749	0.0404	0.0296	0.0640	0.0849
AE+KMeans	0.101	0.0229	0.0339	0.0634	0.0911
DEC	0.098	0.0141	0.0291	0.0632	0.0851
IDEC	0.108	0.0173	0.0280	0.0523	0.0863

Removing `STU_IDCLASS` materially changes individual membership: ARI ranges from 0.098 for DEC to 0.839 for agglomerative clustering, and is approximately 0.10 for all three autoencoder-based methods. Consequently, the profile-specific memberships in the primary analysis are not robust to this feature and should not be treated as stable student classifications. The principal comparative conclusion nevertheless remains: common-space separation is weak for every method (Silhouette 0.0141–0.0867), and DEC and IDEC do not outperform the simpler baselines. Held-out associations also remain modest (reading $\eta^2 = 0.0280$–0.0653, HRL $\eta^2 = 0.0523$–0.1049, and home-language $V = 0.0642$–0.0911). This analysis resolves whether the substantive diagnosis depends on the identifier, but it also strengthens the limitation on interpreting the original five profiles or individual assignments.

8. Conclusion and Future Work

This study compared classical and deep clustering methods as exploratory tools for the Moroccan PIRLS 2021 data. DEC and IDEC produced strongly separated latent representations, but their labels remained weakly separated when every method was evaluated in the common standardized input space. Their high latent-space indices must therefore be interpreted as properties of optimized representations rather than evidence of five naturally distinct learner populations. PCA+KMeans was more stable and achieved better common-space indices than the deep methods, although its own separation remained weak. Dimensionality reduction thus contributes to the observed differentiation, but the results do not establish the superiority or necessity of complex deep clustering objectives for this dataset.

External validation showed that the partitions retain some educational information. Held-out reading achievement and home resources differed across profiles, whereas home-language frequency had a smaller association. Nevertheless, the effect sizes were modest, assignments were unstable across random seeds, and individual memberships were sensitive to missing-data treatment and to exclusion of the identifier-like `IDCLASS` field. The profiles should consequently be viewed as model-dependent summaries of overlapping educational gradients, not stable learner types, causal categories, or a basis for individual educational decisions.

The replication in France, Singapore, and South Africa reproduced the central contrast between strong latent separation and weak common-space separation, although the size and ordering of the educational effects varied

across countries. The ablation analysis also showed that latent dimension and network architecture can materially change cluster membership. In addition, the CPU benchmark found that DEC and IDEC required substantially more computation than KMeans. Taken together, these findings do not demonstrate a clear practical advantage of deep clustering over simpler alternatives, but they support its cautious use for exploratory hypothesis generation.

The study remains limited by same-sample external-variable validation, the absence of expert adjudication and leakage-free downstream prediction, cross-device sensitivity of the neural partitions, non-bitwise-deterministic CUDA latent coordinates, and the lack of held-out-country transfer. In addition, the archived clustering matrix retained several identification, weighting, and sampling-design variables that may have influenced the resulting partitions independently of the substantive educational characteristics. Future work should compare discrete clustering with continuous factor models, incorporate survey-aware imputation and inference, align profiles across countries, and test whether cluster information improves genuinely out-of-sample prediction before the profiles are considered actionable for educational policy.

Data and Code Availability

The complete experimental code, notebooks, aggregate results, environment specification, and data-access instructions are publicly available at https://github.com/RimElmaatoui/PIRLS-2021-Clustering-Reproducibility. The PIRLS 2021 International Database, Edition 2, can be obtained from the official PIRLS 2021 data portal and is identified by the dataset DOI 10.58150/PIRLS_2021_edition_2_including_Log-file_data. The original data files are not redistributed in the GitHub repository; users must obtain them from the IEA under the applicable terms and follow the data-placement instructions provided in the repository.

REFERENCES

1. Abdennour Mohamed Jalil, Imad Hafidi, Lamiae Alami, and ENSA Khouribga. Comparative study of clustering algorithms in text mining context. *IJIMAI*, 3(7):42–45, 2016.
2. Abdennour Jalil, Noureddine Aboutabit, and Imad Hafidi. Synonym reduction in a comparative study of clustering algorithms in text mining. In *International Conference on Information Technology and Communication Systems*, pages 24–34. Springer, 2017.
3. Meriem Amnay, Abdelfattah Toulaoui, Mourad Jabrane, and Imad Hafidi. Unveiling the effectiveness of contextual embeddings in sentiment analysis on moroccan darija: A comparative study with traditional techniques. *IAENG International Journal of Computer Science*, 52(10), 2025.
4. Siham El Bahy, Noureddine Aboutabit, and Imad Hafidi. Analysis and prediction of personality traits using a self-generated database of moroccan instagram users: Impact of gender on image content and quantity on prediction accuracy. *Multimedia Tools and Applications*, 83(29):73583–73606, 2024.
5. Mouad Jbel, Imad Hafidi, and Abdelmoutalib Metrane. Myc: A moroccan corpus for sentiment analysis. In *International Conference of Machine Learning and Computer Science Applications*, pages 59–68. Springer, 2022.
6. Mohamed Hassan Oukhouya, Nora Angour, Noureddine Aboutabit, and Imad Hafidi. Comparative analysis of ardl, lstm, and xgboost models for forecasting the moroccan stock market during the covid-19 pandemic. *Informatica*, 49(14), March 2025.
7. Amal Azeroual, Naoual Attaoui, Mourad Azhari, Soufiane El Adraoui, Benayad Nsiri, and Mohamed Essaaidi. Machine learning in education: An overview of applications and datasets. *Statistics, Optimization & Information Computing*, 16(3):2581–2598, 2026.
8. Zakaria Khoudi, Mourad Nachaoui, and Soufiane Lyaqini. Finding the contextual impacts on students' mathematical performance using a machine learning-based approach. *Infocommunications journal*, 16(Special Issue):12–21, 2024.
9. Zakaria Khoudi, Mourad Nachaoui, and Soufiane Lyaqini. Identifying the contextual factors related to the reading performance of moroccan fourth-grade students from a machine learning-based approach. *Education and Information Technologies*, 29(3):3047–3073, June 2023.
10. Zakaria Khoudi, Nasereddine Hafidi, Mourad Nachaoui, and Soufiane Lyaqini. New approach to enhancing student performance prediction using machine learning techniques and clickstream data in virtual learning environments. *SN Computer Science*, 6(2):139, 2025.
11. Ali Satty and Zakariya M. S. Mohammed. Integrating statistical clustering methods and machine learning to uncover latent population profiles in childhood disability. *Statistics, Optimization & Information Computing*, 16(3):2750–2761, 2026.
12. Ina V. S. Mullis, Matthias von Davier, Pierre Foy, Bethany Fishbein, Katherine A. Reynolds, and Erin Wry. *PIRLS 2021 International Results in Reading*. Boston College, TIMSS & PIRLS International Study Center, 2023.
13. Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8):1295, August 2020.
14. Bao Chong. K-means clustering algorithm: A brief review. *Academic Journal of Computing & Information Science*, 4(5):37–40, 2021.

15. Eric K. Tokuda, Cesar H. Comin, and Luciano da F. Costa. Revisiting agglomerative clustering. *Physica A: Statistical Mechanics and its Applications*, 585:126433, January 2022.
16. Yi Zhang, Miaomiao Li, Siwei Wang, Sisi Dai, Lei Luo, En Zhu, Huiying Xu, Xinzhong Zhu, Chaoyun Yao, and Haoran Zhou. Gaussian mixture model clustering with incomplete data. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 17(1s):1–14, January 2021.
17. Sheng Zhou, Hongjia Xu, Zhuonan Zheng, Jiawei Chen, Zhao Li, Jiajun Bu, Jia Wu, Xin Wang, Wenwu Zhu, and Martin Ester. A comprehensive survey on deep clustering: Taxonomy, challenges, and future directions. *ACM Computing Surveys*, 57(3):1–38, November 2024.
18. Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016.
19. Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. Improved deep embedded clustering with local structure preservation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, IJCAI-2017, page 1753–1759. International Joint Conferences on Artificial Intelligence Organization, August 2017.
20. Matthias von Davier, Ina V. S. Mullis, Bethany Fishbein, and Pierre Foy, editors. *Methods and Procedures: PIRLS 2021 Technical Report*. Boston College, TIMSS & PIRLS International Study Center, 2023.
21. Nina Valentika, I Made Sumertajaya, Aji Hamim Wigena, and Farit Mochamad Afendi. Robust PCA-based multivariate time series clustering for weekly indonesian food commodity prices with missing data. *Statistics, Optimization & Information Computing*, 16(3):1968–1983, 2026.
22. Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.
23. Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.
24. David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.
25. William H. Kruskal and W. Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952.
26. Harald Cramér. *Mathematical Methods of Statistics*, volume 9 of *Princeton Mathematical Series*. Princeton University Press, Princeton, NJ, 1946.