A Reproducible Unsupervised Learning Workflow for Hippocratic Oath Profile Identification Using an Empirically Informed Synthetic Benchmark

Alberto Hananel[1,*,†], Hugo Calienes[2,†]

[1]Universidad Catolica Santo Toribio de Mogrovejo, Chiclayo, Peru
[2]Universidad Peruana Cayetano Heredia, Lima, Peru

Abstract

This paper presents a reproducible and stability-aware workflow for exploratory clustering of ordinal questionnaire data. The workflow is evaluated on an empirically informed synthetic benchmark containing $n^* = 150$ observations and six ordinal analytical variables concerning bioethical education and perceptions related to the Hippocratic Oath. The methodological study was developed within an ongoing doctoral research program in which qualitative interviews informed the construction of an original 10-item Likert questionnaire. The associated empirical quantitative dataset comprises $n_0 = 38$ complete questionnaire responses. The synthetic benchmark is used here as a computational test bed for developing and auditing the multivariate workflow; its $n^* = 150$ observations are not interpreted as additional empirical participants or as increasing the inferential power of the doctoral sample.

Standardized Pearson-based Principal Component Analysis (PCA) is used as the primary low-dimensional representation and is complemented by a polychoric-correlation sensitivity analysis. The first two Pearson components explain 61.94% and 29.54% of standardized variance, respectively (91.49% cumulatively), and both satisfy the eigenvalue-greater-than-one retention criterion. Cluster number is evaluated using Within-Cluster Sum of Squares (WSS), average silhouette width, and the Gap statistic, with all three criteria selecting $k = 3$. The resulting three-cluster solution has balanced sizes of 50/50/50, a mean silhouette width of 0.85470, and perfect agreement between k-means and Ward clustering (ARI = 1.00000). The same partition is recovered across 100 random seeds, leave-one-out perturbations, cluster-wise bootstrap stability assessment, PAM, and the polychoric sensitivity representation. A four-cluster solution, examined as a sensitivity analysis, shows a lower mean silhouette width (0.74379), a k-means–Ward ARI of 0.97879, and a smallest cluster of 16 observations.

Because the results are obtained from a synthetic methodological benchmark, the recovered profiles are not interpreted as empirical subgroups of the doctoral participants or of the broader medical-student population. The contribution is methodological: an auditable workflow that integrates dimensionality reduction, cluster-number selection, cross-algorithm agreement, perturbation stability, measurement-scale sensitivity, and substantive characterization, providing a prespecified analytical framework for subsequent application to empirical questionnaire data.

Keywords

unsupervised learning, cluster stability, Principal Component Analysis, ordinal questionnaire data, synthetic data, Hippocratic Oath

1. Introduction

Exploratory clustering is frequently used to investigate whether multivariate questionnaire responses contain groups of observations with similar response patterns. An interpretable low-dimensional projection, however, does not by itself establish a discrete cluster structure. Principal component analysis (PCA) and clustering methods such as k-means optimize different criteria, and the existence, number, and stability of candidate clusters therefore require explicit assessment rather than visual inference from a low-dimensional representation [1, 2, 3].

IWSAI 2026: 1st International Workshop on Systems, Applications, Artificial Intelligence and Innovation, September 16–18, 2026, Piura, Peru
*Corresponding author.
†These authors contributed equally.
ahananel@usat.edu.pe (A. Hananel); hugo.calienes@upch.pe (H. Calienes)
0000-0003-4113-8623 (A. Hananel); 0009-0001-5899-4365 (H. Calienes)

A second challenge is reproducibility and stability. Because k-means may depend on initialization, multiple starts and a fixed random seed improve computational repeatability, but they do not by themselves establish that a partition is robust to alternative initializations or perturbations of the data. Stability-based approaches consequently evaluate whether comparable partitions are recovered under resampling or related perturbation schemes [4, 5]. Agreement across clustering algorithms provides an additional diagnostic, although such agreement should be interpreted jointly with cluster-number and stability evidence rather than as independent proof of a true underlying partition.

A third issue concerns measurement scale. Likert-type questionnaire items are ordinal, whereas ordinary PCA applied to numeric item scores is based on Pearson associations. Polychoric correlations provide an alternative association representation designed for ordered categorical measurements and may materially affect factorial representations [6]. The present study therefore uses Pearson-PCA as the primary operational representation and explicitly examines whether the downstream clustering solution changes when a polychoric-correlation representation is used.

Pilot questionnaire studies also frequently provide only modest empirical sample sizes. The adequacy of a sample for factor or component recovery cannot be determined from a universal minimum N or a single subject-to-variable ratio. Simulation studies have shown that recovery and stability depend on characteristics such as component saturation, communalities, and factor overdetermination [7, 8, 9]. Exploratory factor solutions may, under favorable conditions, be recovered even with samples below $N = 50$, although such small samples require particularly cautious interpretation and explicit assessment of stability [10]. Resampling approaches have likewise been used to examine the sensitivity of PCA eigenstructure in small empirical datasets [11].

These considerations are particularly relevant to the broader doctoral research context from which the present methodological study emerged. The doctoral investigation combines qualitative interviews with a quantitative questionnaire concerning bioethics education and perceptions related to the Hippocratic Oath. The empirical quantitative dataset contains $n_0 = 38$ complete responses to an original 10-item Likert questionnaire.

The present paper does not use this modest empirical sample to claim population-level factorial or clustering results. Instead, it develops and audits the intended multivariate analytical workflow on a moderate-sized synthetic benchmark ($n^* = 150$) comprising six ordinal analytical variables conceptually informed by the same substantive research domain. The benchmark size is therefore a computational property of the methodological test bed and does not increase the empirical sample size or inferential power of the doctoral study.

Within this setting, the objective is to establish an auditable and stability-aware workflow in which (i) component retention is explicitly justified, (ii) the number of clusters follows predefined complementary criteria rather than a manually imposed value, (iii) agreement across alternative clustering algorithms is quantified, (iv) initialization and perturbation stability are explicitly evaluated, and (v) sensitivity to the treatment of ordinal questionnaire responses is examined before substantive interpretation.

The contribution is therefore methodological integration and computational reproducibility rather than a new clustering algorithm or evidence of population subtypes. The methodological question addressed is whether a candidate clustering of an empirically informed synthetic ordinal-data benchmark can be supported consistently by dimensionality diagnostics, complementary cluster-number criteria, alternative clustering algorithms, perturbation-based stability analyses, and sensitivity to the association representation. The resulting workflow provides a prespecified methodological framework for subsequent application to empirical questionnaire data while deliberately separating within-benchmark robustness from population-level validity.

2. Related Work

K-means partitions observations by minimizing within-cluster squared distances [12], whereas Ward's hierarchical method builds an agglomerative hierarchy by minimizing increases in within-group dispersion [13]. Agreement between such different procedures can be informative, but agreement alone does

not determine the correct value of k.

The silhouette coefficient evaluates how well each observation fits its assigned cluster relative to neighboring clusters and can be summarized by the mean silhouette width [2]. The Gap statistic compares observed within-cluster dispersion against a reference distribution and provides a formal rule for choosing k [3].

Cluster stability adds a complementary dimension. Ben-Hur et al. evaluate the persistence of structure under perturbation [4], while Hennig's cluster-wise bootstrap approach uses Jaccard similarity to identify clusters that are less stable than the overall partition [5]. This is particularly important when a candidate solution produces a small additional cluster.

Partition agreement is quantified here with the Hubert–Arabie Adjusted Rand Index (ARI), which adjusts agreement for chance [14]. Because ARI is a comparison between particular partitions and has an upper bound of 1 for identical assignments, this paper reports the numerical values and the partitions being compared rather than assigning a universal qualitative category to a given ARI value [15].

In contrast to clustering applications that report a single partition obtained from a predefined number of groups or rely primarily on one clustering criterion, the present workflow combines several levels of validation within the same auditable analysis. Specifically, component retention is defined before clustering, the number of clusters is selected through complementary WSS, silhouette, and Gap diagnostics, the resulting partition is compared across k-means, Ward, and PAM, and its stability is examined under alternative random seeds, leave-one-out perturbations, and cluster-wise bootstrap resampling. In addition, the sensitivity of the partition to the ordinal nature of the questionnaire items is evaluated through a polychoric-correlation representation. The contribution therefore lies not in any individual clustering technique, but in integrating these diagnostic, stability, and measurement-sensitivity checks into a single reproducible workflow for questionnaire data.

Recent applications in medical education provide a particularly relevant comparison because they use questionnaire responses to identify attitudinal profiles among medical students. Richelle et al. [16] applied a two-stage procedure combining Ward hierarchical clustering and k-means to Likert-type responses concerning medical students' attitudes, retaining four clusters using Calinski–Harabasz and Duda–Hart criteria. Similarly, Allam et al. [17] applied k-means to medical students' attitude and perception questionnaire domains and selected the number of clusters using the maximum silhouette criterion. Their three analyses selected $k = 2$, with reported silhouette values of 0.325, 0.42, and 0.50.

More recently, Wang et al. [18] used k-means to identify profiles from Likert-type measures of academic attitude, ethical climate, academic pressure, institutional environment, and misconduct in medical education. Their three-cluster solution was selected using both the WSS elbow and silhouette criteria, with multiple random starts.

These studies illustrate that person-centered clustering of perception and attitude questionnaires is already established in medical education, but the validation strategies differ across applications. The present workflow extends this practice by jointly evaluating WSS, silhouette, and Gap criteria, comparing k-means with Ward and PAM, assessing stability under alternative seeds, leave-one-out perturbation and cluster-wise bootstrap resampling, and evaluating sensitivity to Pearson versus polychoric representations of the ordinal items.

3. Data and Methods

3.1. Empirical provenance, synthetic benchmark, and doctoral scope

The present methodological study was developed within an ongoing doctoral research project on bioethics education and perceptions related to the Hippocratic Oath. The empirical component of that project follows a mixed-methods sequence in which qualitative interviews provide the conceptual basis for a quantitative questionnaire. The original quantitative instrument comprises 10 Likert-type items addressing the content of bioethics education, teaching methodology and its impact, and alignment with principles related to the Hippocratic Oath. The empirical quantitative dataset currently comprises $n_0 = 38$ complete questionnaire responses.

The computational dataset analyzed in the present paper is a fixed methodological benchmark of $n^* = 150$ records originally developed for multivariate exploration within the same research line. It contains six ordinal analytical variables: *Compromiso*, *Respeto_Pacientes*, *Valor_Etico*, *Reflexion*, *Aplicacion_Juramento*, and *Valoracion_Codigo*. These variables operationalize conceptually relevant aspects of attitudes toward the Hippocratic Oath, including personal commitment, respect for patients, ethical value, ethical reflection, perceived applicability of the Oath, and the relative valuation of contemporary professional codes.

The six benchmark variables are conceptually connected to the broader qualitative and quantitative framework of the doctoral research, but they should not be interpreted as a one-to-one subset or direct transformation of six items from the current 10-item empirical questionnaire. Likewise, the $n^* = 150$ benchmark observations are distinct from the $n_0 = 38$ complete empirical questionnaire responses and are not presented as 150 independent participants from the doctoral study.

The target size $n^* = 150$ provides a moderate-sized computational environment in which dimensionality reduction, cluster-number selection, alternative clustering algorithms, and perturbation-based stability procedures can be developed and audited. It is not presented as a universal sample-size recommendation for PCA or clustering, nor as a mechanism for increasing the inferential power of the $n_0 = 38$ empirical doctoral sample.

Importantly, no cluster labels, prespecified values of k, centroids, PCA coordinates, or target cluster proportions are used as inputs to the analytical workflow evaluated in this paper. Cluster discovery is performed only after the fixed benchmark has been defined. Accordingly, the resulting profiles are interpreted as computational response patterns within this benchmark rather than as validated population subgroups.

The methodological relationship between the benchmark and the doctoral study is therefore prospective rather than inferential. The benchmark is used to develop, test, and document the sequence of multivariate analytical decisions. When this workflow is applied to the $n_0 = 38$ empirical questionnaire responses, component retention, cluster number, separation, and stability must be reassessed directly from the observed data rather than inherited from the synthetic benchmark.

3.2. Pearson PCA and formal retention rule

The six synthetic ordinal analytical variables are treated as numeric ordered scores and standardized to unit variance.

Component retention is determined before clustering using Kaiser's eigenvalue-greater-than-one rule [19]. Cumulative explained variance and the scree pattern are reported as corroborating diagnostics rather than used to override this predefined retention rule.

The eigenvalues are 3.7166, 1.7726, 0.2228, 0.1195, 0.0898, and 0.0787. Only PC1 and PC2 have eigenvalues greater than one. They explain 61.944% and 29.543% of the standardized variance, respectively, accounting for 91.487% cumulatively (Table 1). Accordingly, the individual scores on PC1 and PC2 define the primary two-dimensional space used for the subsequent clustering analyses. This representation retains 91.487% of the standardized variance while providing a low-dimensional space in which the clustering diagnostics and resulting partitions can be visually audited.

Table 1
Eigenvalues and variance explained by the standardized Pearson PCA.

Component	Eigenvalue	Variance (%)	Cumulative (%)
PC1	3.7166	61.944	61.944
PC2	1.7726	29.543	91.487
PC3	0.2228	3.713	95.200
PC4	0.1195	1.992	97.192
PC5	0.0898	1.496	98.689
PC6	0.0787	1.311	100.000

3.3. Variable coordinates and ordinal sensitivity

Table 2 and Figure 1 report the coordinates (correlation loadings) of the six synthetic analytical variables on the two retained Pearson principal components.

PC1 is characterized by large negative coordinates for `Compromiso` and `Aplicacion_Juramento` and large positive coordinates for `Reflexion`, `Respeto_Pacientes`, and `Valor_Etico`. PC2 is characterized by a large negative coordinate for `Valoracion_Codigo` and positive coordinates for `Valor_Etico` and `Aplicacion_Juramento`. Because the signs of principal components are mathematically arbitrary, these directions should be interpreted through the relative configuration of the variables rather than through the signs themselves. Moreover, the components describe the geometry of the fixed synthetic benchmark and are not interpreted as validated population-level latent constructs.

Because the six synthetic analytical variables are ordinal, an additional sensitivity analysis is performed using a polychoric correlation matrix. The first two eigenvectors of this matrix are used to obtain an alternative two-dimensional projection of the standardized observed scores. This representation is deliberately treated as an ordinal sensitivity projection rather than as an estimate of latent individual factor scores.

Table 2
Variable coordinates (correlation loadings) on the first two standardized Pearson principal components.

Variable	PC1	PC2
Compromiso	-0.949	0.102
Respeto_Pacientes	0.837	0.495
Valor_Etico	0.667	0.710
Reflexion	0.952	-0.016
Aplicacion_Juramento	-0.753	0.528
Valoracion_Codigo	0.443	-0.857

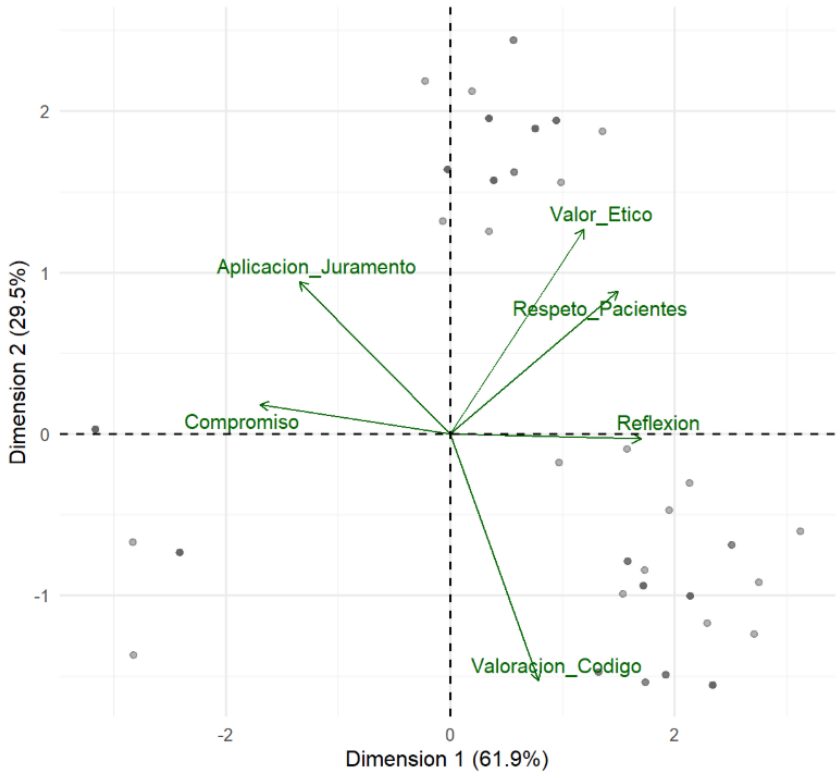

Figure 1: Variable coordinates (correlation loadings) of the six synthetic analytical variables on the two retained standardized Pearson principal components.

The same k-means procedure is then applied to this alternative representation, and the resulting partition is compared with the primary Pearson-PCA partition using the Adjusted Rand Index (ARI). This comparison evaluates whether the clustering result is sensitive to the association structure used to represent the ordinal items.

3.4. Cluster-number selection

Candidate values $k = 2, \ldots, 9$ are evaluated before interpreting the resulting cluster partitions. Three complementary criteria are used: a discrete-curvature elbow diagnostic for the within-cluster sum of squares (WSS), the value of k that maximizes the mean silhouette width [2], and the one-standard-error selection rule for the Gap statistic [3]. Gap estimation is based on 200 Monte Carlo reference samples.

The three criteria are evaluated independently rather than imposing a prespecified number of clusters. In the present benchmark, all three converge on $k = 3$ (Table 3; Figure 2). Accordingly, $k = 3$ is adopted as the primary solution. The $k = 4$ partition is retained only as a sensitivity analysis because it corresponds to the solution reported in the previous version of the manuscript and therefore permits a direct assessment of the consequences of imposing an additional cluster.

Table 3
Cluster-number selection and comparison of the primary $k = 3$ solution with the $k = 4$ sensitivity solution.

Diagnostic/metric	$k = 3$	$k = 4$
WSS discrete elbow	selected	not selected
Silhouette criterion	selected	not selected
Gap one-SE rule	selected	not selected
Mean silhouette	0.85470	0.74379
k-means–Ward ARI	1.00000	0.97879
Cluster sizes	50/50/50	50/50/34/16
Analytical role	primary	sensitivity

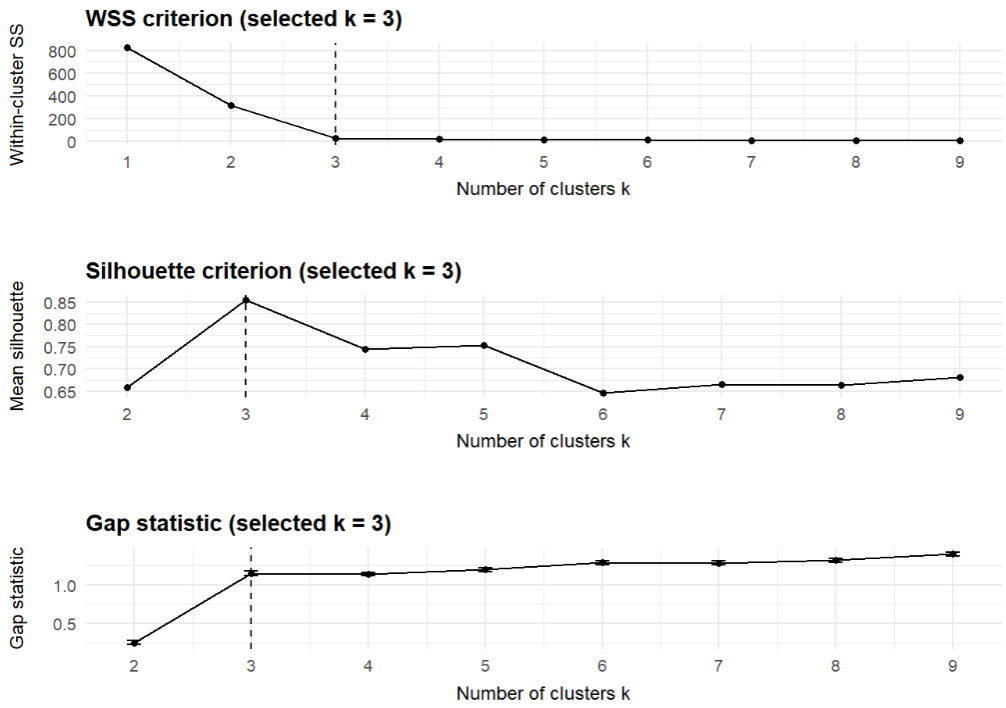

Figure 2: Selection of the number of clusters using WSS, average silhouette width, and the Gap statistic. All three diagnostics select $k = 3$; the dashed vertical line in each panel indicates the value selected by the corresponding criterion.

3.5. Clustering, agreement, and stability

For the primary solution, k-means is run with seed 2026 and `nstart=50`. Ward hierarchical clustering with Euclidean distance and the `ward.D2` criterion is fitted in the same two-component space and cut at the selected value of k. Agreement between the two partitions is quantified using the Adjusted Rand Index (ARI) [14].

Stability is evaluated beyond a single random seed. k-means is refitted for 100 independent seeds (2026–2125), each with 50 starts, and each resulting partition is compared with the reference partition using ARI. In addition, leave-one-out perturbation is used to assess the effect of removing individual observations and refitting the model. Cluster-wise bootstrap stability is evaluated using 200 bootstrap replications and mean Jaccard similarity following Hennig's framework [5].

As complementary sensitivity analyses, Partitioning Around Medoids (PAM) is fitted using the same selected value of k, and its partition is compared with the k-means solution using ARI. The polychoric-PCA representation is likewise clustered using the selected k, allowing the sensitivity of the partition to the ordinal nature of the analytical variables

Finally, a four-cluster solution is fitted as a secondary sensitivity analysis to examine the alternative partition reported in the previous version of the manuscript. The $k = 4$ solution is evaluated using cluster sizes, mean silhouette width, k-means–Ward agreement, cluster-wise bootstrap stability, and a robust-distance diagnostic for the smallest resulting group. This analysis is not used to override the cluster-number selection criteria, but to determine whether the additional cluster represents a stable subdivision of the primary $k = 3$ structure or an isolated set of observations.

3.6. Overall methodological workflow

Figure 3 summarizes the complete analytical workflow. The procedure begins with the fixed synthetic benchmark and proceeds through standardization and dimensionality reduction using the two Pearson principal components retained by the predefined eigenvalue-greater-than-one criterion. The number of clusters is then selected independently using WSS, average silhouette width, and the Gap statistic before fitting the primary clustering solutions.

The selected partition is subsequently subjected to two complementary levels of validation. First, cross-algorithm agreement is evaluated by comparing k-means and Ward hierarchical clustering using the Adjusted Rand Index (ARI). Second, stability is examined under alternative random seeds, leave-one-out perturbations, and cluster-wise bootstrap resampling. Additional sensitivity analyses compare the primary partition with PAM, an ordinal polychoric-correlation representation, and the alternative $k = 4$ solution retained from the previous specification.

This sequential design deliberately separates cluster-number selection, algorithmic agreement, initialization and perturbation stability, and measurement-scale sensitivity. The final stage therefore characterizes response patterns within the fixed synthetic benchmark rather than inferring validated population-level latent profiles.

The empirical questionnaire forms part of a broader mixed-methods doctoral research context in which perceptions surrounding bioethics education and the Hippocratic Oath are investigated through qualitative, documentary, and quantitative evidence. The quantitative component examined here is deliberately exploratory and methodologically embedded: its role is to evaluate whether reproducible multivariate response patterns can generate hypotheses that may subsequently be contrasted with qualitative themes, rather than to replace or statistically generalize the qualitative findings [20].

Accordingly, the focus of the present paper remains the computational workflow—dimensionality reduction, unsupervised profile discovery, stability assessment, and reproducibility—rather than substantive inference about medical ethics.

Data and Code Availability

The individual-level empirical questionnaire data ($n_0 = 38$) belong to the ongoing doctoral research project and are not publicly released because the applicable data-governance and consent conditions do

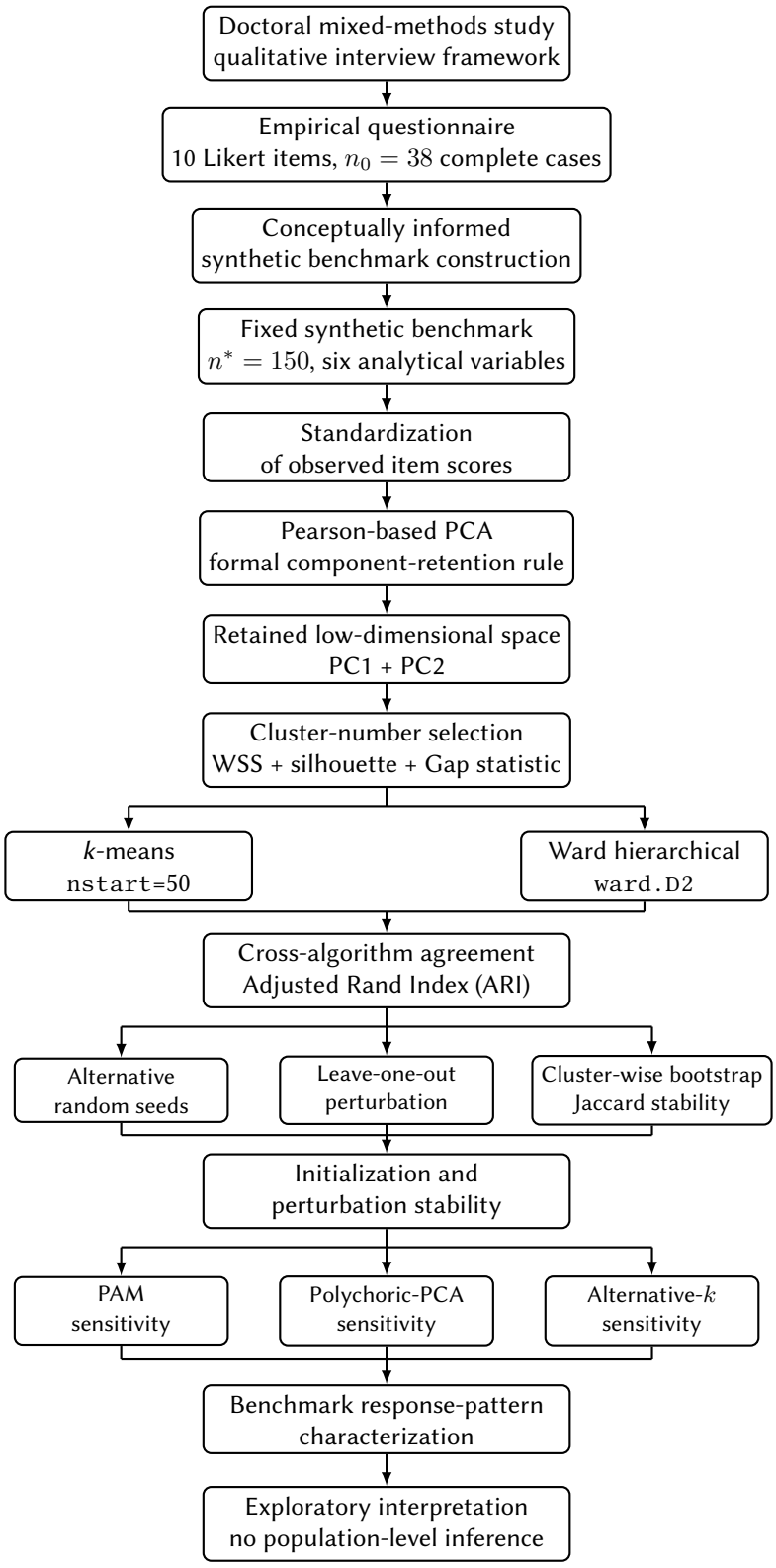

Figure 3: Stability-aware methodological workflow developed within the doctoral research framework. The empirical study comprises $n_0 = 38$ complete responses to a 10-item Likert questionnaire, while a conceptually informed synthetic benchmark ($n^* = 150$, six analytical variables) is used to develop and audit the multivariate workflow.

not authorize unrestricted dissemination of participant-level responses.

To support computational reproducibility, the fixed synthetic benchmark analyzed in the present methodological study and the complete R scripts for PCA, cluster-number selection, clustering, stability assessment, and sensitivity analyses are provided as supplementary research materials. The synthetic benchmark is a methodological test bed and should not be interpreted as a public-use replica of the confidential empirical dataset.

4. Results

4.1. Primary three-cluster solution

The cluster-selection procedure yielded $k = 3$. The k-means solution contains three equal groups of 50 observations, with a mean silhouette width of 0.85470.

Figure 4 shows the primary partition in the retained Pearson-PCA space, whereas Figure 5 shows the corresponding Ward hierarchical solution.

The k-means–Ward ARI is 1.00000, indicating identical partitions up to label permutation. Across 100 alternative seeds, both the median and minimum ARI relative to the canonical k-means partition are 1.00000. The minimum leave-one-out ARI is also 1.00000, all three cluster-wise bootstrap mean Jaccard values are 1.00000, and the k-means–PAM ARI is 1.00000. The Pearson-PCA versus polychoric-sensitivity clustering ARI is likewise 1.00000 (Table 4).

These results indicate complete computational persistence of the three-cluster solution under the sensitivity analyses considered on this fixed synthetic benchmark; they do not establish external validity.

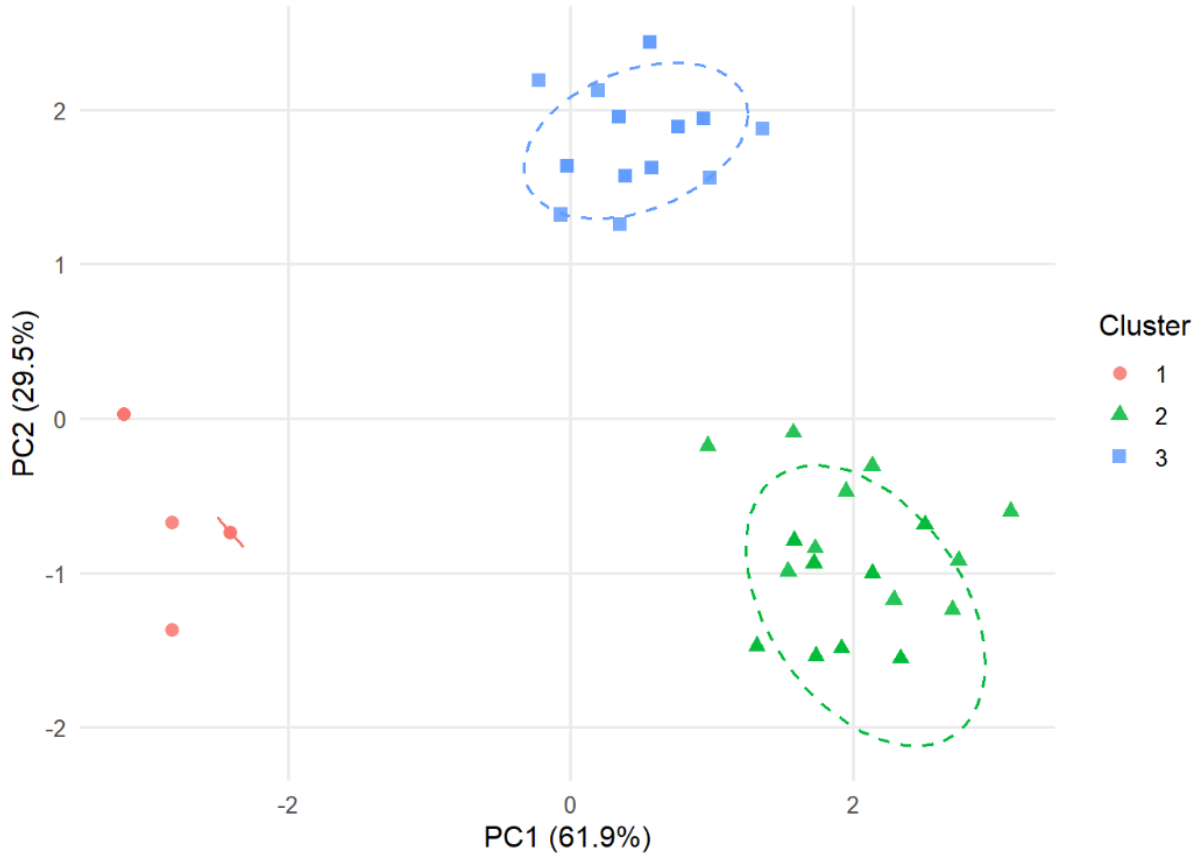

Figure 4: Primary k-means partition for $k = 3$ in the retained Pearson-PCA space. Cluster numbers are computational labels and do not represent real population segments.

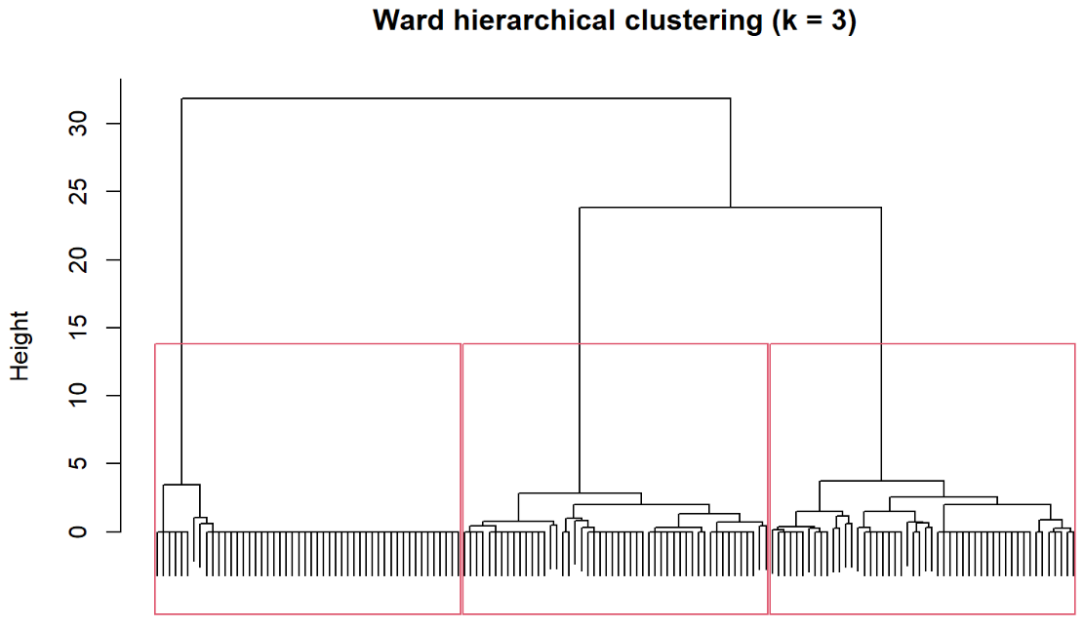

Figure 5: Ward hierarchical clustering in the same two-component space, cut at $k = 3$.

Table 4
Stability and sensitivity diagnostics for the primary $k = 3$ solution.

Diagnostic	Value
Mean silhouette width	0.85470
k-means–Ward ARI	1.00000
Minimum ARI across 100 seeds	1.00000
Median ARI across 100 seeds	1.00000
Minimum leave-one-out ARI	1.00000
Minimum cluster-wise bootstrap Jaccard	1.00000
k-means–PAM ARI	1.00000
Pearson-PCA–polychoric-PCA ARI	1.00000

4.2. Benchmark response patterns

To characterize the three computational clusters in terms of the six synthetic analytical variables, Table 5 reports the mean response for each variable within each cluster.

Cluster 1 is characterized by higher `Compromiso` and lower `Respeto_Pacientes`, `Valor_Etico`, and `Reflexion`. Cluster 2 shows higher `Valoracion_Codigo` and `Reflexion`, together with lower `Compromiso` and `Aplicacion_Juramento`. Cluster 3 shows higher `Respeto_Pacientes` and `Valor_Etico`, together with lower `Valoracion_Codigo`.

These patterns are descriptive summaries of the synthetic records and should not be interpreted as estimated population profiles.

Table 5
Means of the six synthetic analytical variables by primary computational cluster ($k = 3$).

Cluster	Compromiso	Respeto	Valor Etico	Reflexion	Aplicacion	Valoracion Codigo
1	4.02	2.00	1.98	2.00	2.12	3.00
2	1.30	4.12	3.10	3.88	1.12	4.66
3	2.52	4.58	3.96	3.16	2.00	1.92

4.3. Four-cluster sensitivity analysis and substantive interpretation

The $k = 4$ solution is not adopted as the primary statistical result because none of the three cluster-number diagnostics selects four clusters. The three-cluster solution therefore remains the parsimonious and statistically preferred representation of the benchmark data.

Nevertheless, the four-cluster solution provides a potentially informative sensitivity analysis. Its mean silhouette width decreases from 0.85470 to 0.74379 and its k-means–Ward ARI decreases from 1.00000 to 0.97879, while the cluster sizes change from 50/50/50 to 50/50/34/16. Thus, increasing k from three to four does not fragment all three primary groups; rather, it primarily subdivides one of them into two smaller response patterns.

The smallest group contains 16 observations and has a mean silhouette width of 0.45126. Although this provides weaker statistical support for treating the group as an independent cluster, the observations show elevated robust-distance values under the corresponding diagnostic and the minimum cluster-wise bootstrap Jaccard value for the four-cluster solution is 0.7920. Consequently, the additional group should not be interpreted merely as a handful of isolated observations, but neither does the present evidence justify promoting $k = 4$ to the primary solution.

From a substantive perspective, this subdivision may be relevant because attitudes toward the Hippocratic Oath need not vary only between broadly separated profiles. A larger response group may contain different degrees of orientation toward the traditional symbolic and ethical role of the Oath versus contemporary professional codes. The four-cluster solution can therefore be regarded as an exploratory representation of possible within-profile heterogeneity that is concealed by the more parsimonious three-cluster solution.

Because the present observations are synthetic, these patterns must not be interpreted as empirically established groups of physicians or medical students. Rather, the $k = 4$ result generates a substantive hypothesis for subsequent validation with real medical data: a broad attitudinal profile identified under $k = 3$ may contain distinguishable subprofiles reflecting different degrees of adherence to traditional and contemporary frameworks of medical ethics.

5. Discussion

The three-cluster solution is retained as the primary representation because all three cluster-number diagnostics—WSS, silhouette, and Gap—converged on $k = 3$. This solution also showed a higher mean silhouette width than $k = 4$, perfect agreement between k-means and Ward clustering, balanced cluster sizes, and complete persistence under the perturbation analyses considered. Taken together, these results provide convergent evidence that the three-cluster partition is the most stable and parsimonious representation of the structure present in the fixed synthetic benchmark.

These stability results should nevertheless be interpreted conservatively. The Adjusted Rand Index measures agreement between partitions; a value near one does not establish that the corresponding number of clusters is scientifically or clinically correct. Two clustering algorithms may agree strongly even on an over-partitioned solution, so cross-algorithm agreement should be interpreted jointly with cluster-number diagnostics and stability evidence. Likewise, computational stability within a fixed benchmark does not establish external validity or confirm that equivalent profiles occur in the target population.

A central limitation concerns the relationship between the methodological benchmark and the empirical doctoral study. The doctoral quantitative dataset contains $n_0 = 38$ complete questionnaire responses to the original 10-item instrument, whereas the present computational analyses are performed on a fixed synthetic benchmark of $n^* = 150$ observations and six analytical variables. The benchmark was developed to construct and audit the analytical workflow within the same substantive research domain; it is not an enlarged version of the empirical sample and its numerical geometry should not be projected onto the 38 participants.

Consequently, the pronounced low-dimensional geometry observed here—with 91.49% of standardized variance explained by the first two Pearson principal components—and the highly stable three-cluster

partition are properties of the methodological benchmark. They establish that the proposed sequence of dimensionality reduction, cluster-number selection, cross-algorithm comparison, and stability assessment can be implemented and audited coherently. They do not establish that the same component structure or the same three profiles will necessarily be recovered from the empirical doctoral responses.

This distinction is particularly important for subsequent empirical application of the workflow. Component retention must be reassessed from the empirical correlation structure, the number of clusters must be selected again rather than fixed at $k = 3$, and stability must be re-evaluated under perturbation. With a modest empirical sample, failure to reproduce the benchmark geometry would itself constitute an informative empirical result rather than a failure of the methodological framework.

The ordinal-measurement issue was addressed through a dedicated sensitivity analysis. Pearson PCA was retained as the primary operational representation, whereas a polychoric-correlation representation was used to assess the consequences of treating the analytical variables explicitly as ordinal. Both representations yielded the same $k = 3$ partition (ARI = 1.00000). For this benchmark, therefore, the clustering solution was insensitive to the association representation considered. This result should not be generalized to imply that Pearson PCA is preferable for ordinal data; rather, it indicates that the principal clustering conclusion was robust to this particular measurement-level choice.

The magnitude of the separation observed in the present benchmark can be placed in the context of questionnaire-based clustering studies in medical education. Allam et al. [17] used the maximum silhouette criterion to select two clusters in three analyses of medical students' attitudes and perceptions, reporting silhouette values of 0.325, 0.42, and 0.50. The mean silhouette of 0.85470 obtained for the present $k = 3$ solution is numerically higher. This comparison should not be interpreted as evidence that the present workflow or partition is superior, because the studies differ substantially in sample size, questionnaire content, dimensionality, and underlying data geometry. In particular, the present value quantifies separation within an empirically informed synthetic benchmark and should not be interpreted as an estimate of the separation expected in an independent medical-student sample.

Richelle et al. [16], working with Likert-type attitudes of medical students, combined Ward clustering and k-means but retained four clusters using Calinski–Harabasz and Duda–Hart criteria. Wang et al. [18], in an ethics-related medical-education questionnaire, selected three k-means profiles using both the WSS elbow and silhouette criteria. Together, these studies illustrate that the number of profiles recovered from medical-perception questionnaires depends on both the response structure and the validation criteria employed. Within the present benchmark, the convergence of WSS, silhouette, and Gap on $k = 3$, followed by cross-algorithm and perturbation-based stability analyses, provides stronger internal support for the primary partition than reliance on a single cluster-number diagnostic. This constitutes evidence of within-benchmark robustness, however, rather than evidence that three corresponding groups necessarily exist in the underlying population.

The four-cluster solution remains informative as a secondary sensitivity analysis. Although it was not selected by the formal cluster-number diagnostics and showed a lower mean silhouette width, it subdivided one of the primary groups rather than producing a completely different global structure. This suggests that the more parsimonious three-cluster representation may conceal finer within-profile heterogeneity. In the context of medical ethics, such heterogeneity could potentially reflect different degrees of orientation toward the traditional symbolic role of the Hippocratic Oath and toward contemporary professional codes. Because the current evidence is derived from a synthetic methodological benchmark rather than directly from the empirical doctoral dataset, this interpretation should be treated strictly as a hypothesis for subsequent evaluation with empirical medical data, rather than as evidence of clinically or educationally established subgroups.

The possibility of heterogeneous attitudes toward the Oath nevertheless has substantive precedent. Yakir and Glick [21] reported changes in medical students' attitudes toward the physician's oath across stages of training, while Baumgartner and Flores [22] found that, although most surveyed graduating students considered the original Hippocratic Oath relevant, preferences differed regarding its original and adapted forms. These studies did not apply clustering and therefore cannot validate the profiles identified here, but they provide empirical motivation for investigating whether attitudinal heterogeneity surrounding the Oath can be represented as reproducible response profiles in future studies using

independent medical-student data.

Direct numerical comparison with ARI values from unrelated published datasets is not used to claim superiority because ARI depends on the particular partitions and data being compared. Instead, the present study emphasizes controlled same-benchmark comparisons between k-means and Ward, k-means and PAM, Pearson-PCA and polychoric-PCA representations, and the $k = 3$ and $k = 4$ solutions. These comparisons are more appropriate for assessing methodological sensitivity because they hold the benchmark constant while varying specific analytical choices.

A further limitation concerns comparison of the proposed workflow with previously published cluster-selection pipelines. Although the preceding literature provides quantitative and methodological context from questionnaire-based clustering studies in medical education, those studies used different datasets, instruments, sample sizes, and response structures and therefore do not constitute a controlled head-to-head benchmark. Consequently, the relative performance of the present workflow against alternative published pipelines cannot be established from the current analysis. Future work should apply competing cluster-selection workflows—including model-based, density-based, fuzzy, and other clustering approaches—to the same independent empirical datasets and compare them under a common evaluation protocol incorporating cluster-number selection, separation, stability, sensitivity to representation, and computational reproducibility.

The immediate next methodological stage is application of the prespecified workflow to the $n_0 = 38$ complete empirical questionnaire responses available within the doctoral study. This analysis will permit direct assessment of whether reproducible component and response-profile structures emerge from the observed data when cluster number and stability are reassessed de novo. Given the modest empirical sample size, such results must remain exploratory and be accompanied by explicit stability assessment.

A subsequent validation stage should examine any reproducible empirical patterns in a larger independently collected medical-student sample and evaluate their correspondence with themes obtained from the qualitative and documentary components of the doctoral research. Until such external replication is available, the contribution of the present paper remains primarily methodological: it establishes a transparent and stability-aware workflow while explicitly separating computational robustness from claims of population-level validity.

6. Conclusions

The reproducible workflow identifies $k = 3$ as the primary clustering solution. Three independent cluster-number diagnostics support this selection, and the resulting solution has balanced 50/50/50 cluster sizes, a mean silhouette width of 0.85470, and perfect agreement between k-means and Ward clustering (ARI = 1.00000). The same partition is preserved across the implemented random-seed, leave-one-out, bootstrap, PAM, and polychoric-sensitivity analyses.

The $k = 4$ solution is retained as a secondary sensitivity analysis. Although it is not supported as the primary solution by the cluster-number diagnostics and has a lower mean silhouette width, it reveals a subdivision of one of the three primary groups. This result suggests potential within-profile heterogeneity that may be substantively relevant in the context of attitudes toward the Hippocratic Oath and contemporary professional codes. Such heterogeneity constitutes a hypothesis for future evaluation with empirical medical data rather than evidence of an established population subgroup.

The main contribution is methodological. Explicit component-retention rules, cluster-number diagnostics, complementary clustering algorithms, perturbation-based stability assessment, ordinal sensitivity analysis, loading inspection, and transparent data provenance provide an auditable framework for exploratory profile identification without relying on a manually imposed cluster count or an isolated agreement statistic.

The results reported here pertain to a synthetic methodological benchmark and are not estimates of empirical population profiles. Their primary value is to establish a transparent, reproducible, and stability-aware analytical protocol before its application to empirical questionnaire responses. In

that empirical stage, component retention, cluster number, and stability must be reassessed from the observed data rather than inherited from the benchmark. The present workflow therefore provides a methodological bridge between the mixed-methods construction of the doctoral questionnaire and its subsequent multivariate empirical analysis.

Acknowledgments

The authors gratefully acknowledge the institutional support provided by Universidad Catolica Santo Toribio de Mogrovejo (USAT) and Universidad Peruana Cayetano Heredia (UPCH). Special acknowledgment is extended to Dr. Patricia Campos Olazabal, Rector of USAT, for the financial support provided for the publication of research articles derived from this investigation.

This work is derived from the doctoral research project entitled "Análisis multivariante utilizando caras de Chernoff en la percepción sobre la enseñanza de la Ética Médica que tienen internos de medicina de tres universidades públicas peruanas en relación al Juramento Hipocrático" (SIDISI No. 209996), conducted by Hugo Calienes (Universidad Peruana Cayetano Heredia, UPCH) as part of the requirements for the Doctor of Medicine degree at UPCH, under the supervision of Alberto Hananel Baigorria (Universidad Catolica Santo Toribio de Mogrovejo, USAT).

Declaration on Generative AI

During the preparation of this manuscript, the first author used ChatGPT-5 to verify information and assist in improving the clarity and presentation of the text. DeepL Pro was also used exclusively for language translation purposes. All AI-assisted content was subsequently reviewed and edited by the first author as necessary. The authors take full responsibility for the accuracy, integrity, and final content of the manuscript.

References

[1] C. Ding, X. He, K-means clustering via principal component analysis, in: Proceedings of the 21st International Conference on Machine Learning, 2004. doi:10.1145/1015330.1015408.

[2] P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20 (1987) 53–65. doi:10.1016/0377-0427(87)90125-7.

[3] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63 (2001) 411–423. doi:10.1111/1467-9868.00293.

[4] A. Ben-Hur, A. Elisseeff, I. Guyon, A stability based method for discovering structure in clustered data, in: Pacific Symposium on Biocomputing, 2002, pp. 6–17.

[5] C. Hennig, Cluster-wise assessment of cluster stability, Computational Statistics & Data Analysis 52 (2007) 258–271. doi:10.1016/j.csda.2006.11.025.

[6] F. P. Holgado-Tello, S. Chacón-Moscoso, I. Barbero-García, E. Vila-Abad, Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables, Quality & Quantity 44 (2010) 153–166. doi:10.1007/s11135-008-9190-y.

[7] E. Guadagnoli, W. F. Velicer, Relation of sample size to the stability of component patterns, Psychological Bulletin 103 (1988) 265–275. doi:10.1037/0033-2909.103.2.265.

[8] R. C. MacCallum, K. F. Widaman, S. Zhang, S. Hong, Sample size in factor analysis, Psychological Methods 4 (1999) 84–99. doi:10.1037/1082-989X.4.1.84.

[9] D. J. Mundfrom, D. G. Shaw, T. L. Ke, Minimum sample size recommendations for conducting factor analyses, International Journal of Testing 5 (2005) 159–168. doi:10.1207/s15327574ijt0502_4.

[10] J. C. F. de Winter, D. Dodou, P. A. Wieringa, Exploratory factor analysis with small sample sizes, Multivariate Behavioral Research 44 (2009) 147–181. doi:10.1080/00273170902794206.

[11] S. S. Shaukat, T. A. Rao, M. A. Khan, Impact of sample size on principal component analysis ordination of an environmental data set: Effects on eigenstructure, Ekológia (Bratislava) 35 (2016) 173–190. doi:10.1515/eko-2016-0014.

[12] J. A. Hartigan, M. A. Wong, A k-means clustering algorithm, Journal of the Royal Statistical Society: Series C (Applied Statistics) 28 (1979) 100–108. doi:10.2307/2346830.

[13] J. H. Ward, Hierarchical grouping to optimize an objective function, Journal of the American Statistical Association 58 (1963) 236–244. doi:10.1080/01621459.1963.10500845.

[14] L. Hubert, P. Arabie, Comparing partitions, Journal of Classification 2 (1985) 193–218. doi:10.1007/BF01908075.

[15] D. Steinley, Properties of the Hubert-Arabie adjusted rand index, Psychological Methods 9 (2004) 386–396. doi:10.1037/1082-989X.9.3.386.

[16] L. Richelle, M. Dramaix-Wilmet, Q. Vanderhofstadt, C. Kornreich, Cluster analysis of medical students' attitudes regarding people who use drugs: a first step to design a tailored education program, BMC Medical Education 24 (2024) 490. doi:10.1186/s12909-024-05380-8.

[17] A. H. Allam, N. K. Eltewacy, Y. J. Alabdallat, et al., Knowledge, attitude, and perception of arab medical students towards artificial intelligence in medicine and radiology: A multi-national cross-sectional study, European Radiology 34 (2024) 1–14. doi:10.1007/s00330-023-10509-2.

[18] P. Wang, F. Li, J. Li, J. Yang, Institutional environment, academic attitude, and misconduct among medical students: a structural equation modeling and cluster analysis, BMC Medical Education 26 (2026) 1098. doi:10.1186/s12909-026-09442-x.

[19] H. F. Kaiser, The application of electronic computers to factor analysis, Educational and Psychological Measurement 20 (1960) 141–151. doi:10.1177/001316446002000116.

[20] M. D. Fetters, L. A. Curry, J. W. Creswell, Achieving integration in mixed methods designs: Principles and practices, Health Services Research 48 (2013) 2134–2156. doi:10.1111/1475-6773.12117.

[21] A. Yakir, S. M. Glick, Medical students' attitudes to the physician's oath, Medical Education 32 (1998) 133–137.

[22] F. Baumgartner, G. Flores, Contemporary medical students' perceptions of the hippocratic oath, The Linacre Quarterly 85 (2018) 63–73. doi:10.1177/0024363918756389.