Eurasian Journal of Educational Research 118 (2026) 107-125

Eurasian Journal of Educational Research
www.ejer.com.tr

Evaluation of a Physical Education Teacher Program Using the CIPPIEST Model: Validating Learners' Attitudes Based on Knowledge and Skills

Battsetseg Gonchoo[1], Agiimaa Gundalai[2]*, Janchiv Shinebayar[3], Bayartsetseg Nergui[4], Narmandakh Jamba[5]

ARTICLE INFO

Article History:
Received: 01 September 2025
Received in revised form: 11 November 2025
Accepted: 30 December 2025
DOI: 10.14689/ejer.2026.118.07

Keywords
Program evaluation; cluster analysis; learning attitudes; CIPPIEST model; Classical Test Theory.

ABSTRACT

Purpose This study evaluates program learning outcomes in a Physical Education Teacher (PET) program by integrating validated assessment data with cluster-based classification. **Methodology** Using the CIPPIEST program evaluation framework, the product/outcome component was examined through graduation examination results and compulsory course performance data from 67 students who completed the PET program (2020–2024) at the Mongolian National University of Education. Classical test theory was applied to analyze the difficulty, discriminant power, and reliability of randomly selected test items from the graduation examination. **Results** The results indicated acceptable psychometric properties (p = 0.79–0.81; D = 0.54–0.57; r_{pbi} = 0.93–0.96), supporting the reliability of the assessment data. Based on integrated knowledge and skills performance across general, teacher education, professional, and practicum courses, a two-step cluster analysis classified graduates into five distinct learning attitudes. The findings show that 80.7% of graduates were classified within higher learning attitudes clusters, while 19.3% were grouped within comparatively lower learning attitudes. **Implications for research and practice** The study demonstrates that integrated performance patterns in knowledge and skills can serve as an empirical basis for identifying learning attitudes in program evaluation contexts. The results provide evidence-based implications for curriculum revision and targeted professional development within PET.

[1] School of Physical Education, Mongolian National University of Education, Ulaanbaatar, Mongolia
ORCID: https://orcid.org/0009-0000-8940-1078 , Email: g_battsetseg@msue.edu.mn
[2] School of the Humanities and Social Sciences, Mongolian National University of Education, Ulaanbaatar, Mongolia
ORCID: https://orcid.org/0009-0008-0303-1124 , Email: agiimaa@msue.edu.mn
[3] School of Education Study, Mongolian National University of Education, Ulaanbaatar, Mongolia
ORCID: https://orcid.org/0000-0003-1207-3322 , Email: shinebayar@msue.edu.mn
[4] School of the Humanities and Social Sciences, Mongolian National University of Education, Ulaanbaatar, Mongolia
ORCID: https://orcid.org/0009-0002-4034-0045 , Email: bayartsetseg@msue.edu.mn
[5] School of Fine Arts and Technology, Mongolian National University of Education, Ulaanbaatar, Mongolia
ORCID: https://orcid.org/0009-0005-1684-5176 , Email: narmandakh.j@msue.edu.mn
*Correspondence: agiimaa@msue.edu.mn

Introduction

Physical Education Teacher (PET) education programs play a critical role in preparing qualified professionals who possess not only disciplinary knowledge and pedagogical skills but also positive learning attitudes that support lifelong professional development. In recent decades, program evaluation models have been increasingly applied in higher education to assess the quality, effectiveness, and outcomes of teacher education programs. Among these models, the Context–Input–Process–Product (CIPP) framework and its extended version, CIPPIEST (Stufflebeam, 2015), have been widely adopted due to their comprehensive and systematic structure. The CIPP model effectively implemented KTSP physical education learning at Aceh Besar District Public High Schools, resulting in systematic, planned, regular, and continuous improvements in student learning (Maulana, 2024). The implementation of the Independent Curriculum in elementary schools needs a more differentiated and contextual strategy, along with sustainable support for teacher professional development and improving educational infrastructure (Mulyadi et al., 2024). Chu et al. (2022) found that the sports education model significantly improves college students' learning motivation, affective engagement, cognition, and behavior compared to traditional physical education methods. The model also enhances students' responsibility, leadership skills, and active participation, demonstrating its effectiveness as an innovative teaching approach in physical education (Chu et al., 2022). A qualitative case study in Indonesian elementary schools using the CIPP model revealed significant variations in the implementation of the Independent Curriculum for physical education, highlighting the need for differentiated, contextual strategies and sustained support for teacher development and infrastructure improvement (Mulyadi et al., 2024).

The PET program at the Mongolian National University of Education (MNUE), established in 1955, is one of the longest-standing teacher education programs in Mongolia. Since 2014, MNUE has implemented large-scale reforms across all degree levels, resulting in the accreditation of numerous programs by the Mongolian National Commission for Education Accreditation. Building on this reform, the PET program has been evaluated since 2021 using the CIPP and CIPPIEST models, primarily through self-assessment reports and process-oriented evaluations.

A comprehensive self-assessment report for the PET bachelor's program was developed and accredited by the MNCEA on June 20, 2020. This report addressed key areas including program planning, educational activities, the learning environment, student development activities, and quality assurance. In addition, evaluations of the implementation of the PET bachelor's program at the School of Physical Education of MNUE have been conducted along with comparative studies of program implementation (Ganbaatar et al., 2021). Furthermore, the entrance assessment system for new students in the PET bachelor's program was evaluated, and the findings were published in a separate research article.

However, within the Mongolian higher education context, program evaluation has largely emphasized contextual, input, and process-related dimensions, while systematic empirical evaluation of program outcomes remains limited. This limitation has become particularly evident following the initiation of curriculum reform in 2024, which necessitates the revision and alignment of Program Learning Outcomes with contemporary educational and professional standards.

Within the CIPPIEST framework, program outcomes are commonly operationalized through students' knowledge, skills, values, and learning attitudes (Table 1). While learning attitudes are frequently measured using self-report questionnaires, such instruments were not administered to recent graduates of the PET program. Moreover, in practice-oriented fields such as PET, learning attitudes may also be reflected in sustained engagement and performance across theoretical and practical learning contexts.

Table 1

Evaluation Components of the PET program According to the CIPPIEST model

Context	Input	Process	Product	Impact Effectiveness Sustainability Transportability
Mission Goals Objectives	Resources Infrastructure Curriculum Content	Teaching Learning Co-curricular activities	Knowledge Values Attitudes Results	Curriculum efficiency

Research has shown that the correlation and significance between learning styles and students' knowledge, skills, and attitudes is high (Ernest, 2015; Halim et al., 2022). The combined (integrated) influence of university students' learning attitudes, knowledge, and digital skills significantly determines learners' participation in modern educational environments (Şenay, 2013).

Therefore, the present study focuses exclusively on the Product component of the CIPPIEST model by empirically examining graduates' knowledge, skills, and learning attitudes. Specifically, this study aims to (1) validate the assessment of knowledge and skills through graduation examination analysis using Classical Test Theory, and (2) classify students' learning attitudes based on integrated knowledge and skills performance using cluster analysis. By adopting this approach, the study seeks to contribute an alternative method for evaluating learning attitudes and program outcomes in the PET program during periods of curriculum reform.

Literature Review

Numerous studies at the international level have effectively utilized the CIPP model to evaluate Bachelor of Education (BEd) programs. The CIPP Model for program evaluation was developed in the late 1960s as an alternative to objective, testing, and experimental design views (Stufflebeam, 2003). This study investigated the effectiveness of a peer mentoring program at a Spanish university over five academic years (Aitana et al., 2025), Toosi published a systematic review article (Toosi et al., 2021). The Context, Input, Process, and Product (CIPP) model is a comprehensive perspective that attempts to provide information in order to make the best decisions related to CIPP.

Assessing the structural relationships between the CIPP model components in a teacher education program. The CIPP model components directly influence preservice teachers' perceptions of teacher preparation programs, with direct relationships between context, input, process, and product components (Alquraan et al., 2025). Sankaran and Saad (2022) found that while the BEd program in Malaysian polytechnics was effectively structured

using the CIPP model, limitations in infrastructure, lesson planning, and teaching materials hindered its full implementation (Sankaran & Saad, 2022).

The elementary teacher-learner program in physical education is effective in facilitating teacher teaching, learning, and professional development (Suryobroto et al., 2018). The implementation of the independent curriculum model in physical education subjects has been fulfilled, but weakness indicators related to teacher consistency in implementing learning still exist (Wahidah et al., 2023). Evaluation of sports and health Physical Education Program in using the CIPP evaluation model (context, input, process, product) (Setyadi et al., 2022). An evaluation of the instruction, activities, and assessment methods of the bachelor's level Health and Physical Education (HPE) program being implemented in colleges of Khyber Pakhtunkhwa province was conducted within the framework of the CIPP model, and it was concluded that some deficient aspects were identified in the process component, including rehabilitation techniques, the role of mentorship in stimulating students' interest, and the regular arrangement of physical activities, and these areas need to be improved (Iqbal et al., 2022). The CIPP model effectively implemented KTSP physical education learning at Aceh Besar District Public High Schools, resulting in systematic, planned, regular, and continuous improvements in student learning (Maulana, 2024).

The study found a statistically significant correlation between student attitude and academic performance (Bellido-Medina et al., 2023) and a significant positive correlation between students' attitudes toward mathematics and their academic performance (Musa et al., 2022; Umac et al., 2025).

Although research studies related to program evaluation in Mongolia are relatively few, some experiences exist. Implementing educational evaluation using appropriate methodologies and scientifically-based models serves as important evidence for stakeholders to accept evaluation results (Tudevdagva & Sanjdorj, 2022). Furthermore, an evaluation was conducted using the CIPP model methodology on the “Graphic Design” bachelor’s program at MNUE and Inner Mongolia Normal University in China, which demonstrated that this is an effective methodology for identifying the strengths and weaknesses of programs, assessing opportunities and risks, and providing fundamental information for planning program development, formulating policies, and making decisions (Erjing, 2025).

Although numerous studies have applied the CIPP model in teacher education, most of them primarily focus on context, input, and process components, while empirical evaluation of product outcomes-particularly students’ learning attitudes-remains limited.

Methodology

Research Design

This study employed a Program Evaluation Design, integrating Classical Test Theory (CTT) to evaluate the psychometric properties of assessment instruments and cluster analysis to classify students into distinct performance groups, in order to assess the knowledge and skill performance of students in the Physical Education Teacher (PET) program at the Mongolian National University of Education (MNUE) across the 2023–2026 academic years. Guided by the CIPP evaluation model, the study concentrated on the Product component

examining students' learning outcome performance, while also investigating learning attitudes to provide a systematic evaluation of PET program effectiveness.

Data Collection Procedures

Two distinct datasets were extracted from the University Learning Management System (ULMS) of the Mongolian National University of Education (MNUE). These datasets were used to examine whether professional courses have reliably assessed the knowledge and skills of graduates of the PET program.

The first dataset comprised graduation examination results from students who completed the PET bachelor's program in 2024. Item and test analyses were conducted using graduation examination data from a total of 67 students, including 33 graduates from the winter semester and 34 from the spring semester. The graduation examination consisted of approximately 400 closed-ended test items and around 50 open-ended tasks covering content from professional physical education subjects. From this pool, 50 test items were randomly selected for detailed analysis: 41 reading comprehension items were classified as measures of knowledge, and 9 written-response tasks were classified as measures of skills. The examination system was configured to randomly assign identical tests to all students within a single examination session. The difficulty index, discriminant power, and reliability of the selected test items were analyzed.

The second dataset consisted of grades obtained from compulsory courses. The design used for collecting this dataset is illustrated in Figure 1. The study focused on students who enrolled in the PET program in 2020 and graduated in 2024 (n = 67). Students were classified according to their attitudes toward learning based on assessments of their knowledge (examination and test scores) and skills (assignment scores) in compulsory courses, including general education, teacher education, professional, and practicum courses. A two-step cluster analysis method was employed for this classification. (Figure 1)

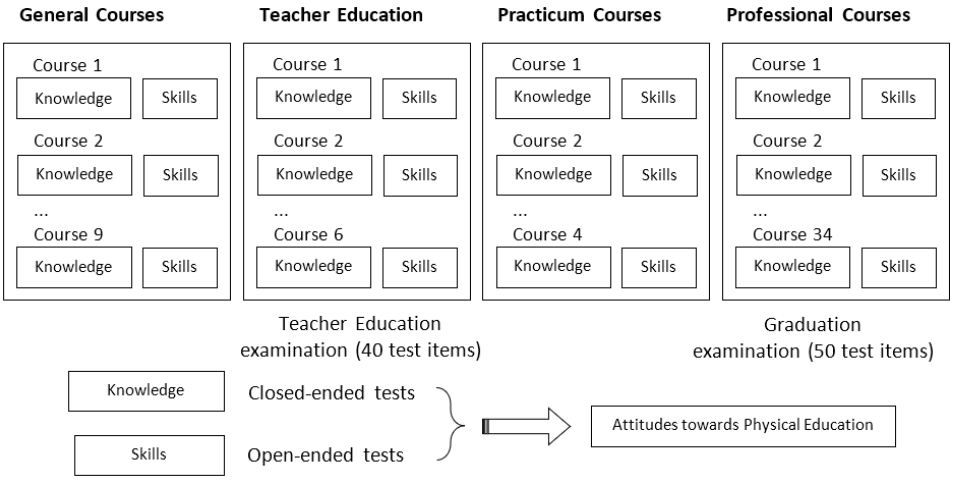


Figure 1: Design for data collection

Data Analysis Methods

Tests are used to assess the knowledge and skills of general education school students and university students. Multiple-choice, matching, fill-in-the-blank, written response, and numerical answer test types are predominantly used. Among these, multiple-choice tests constitute the majority. Analysing test data has become an important research topic for teachers and education researchers (Himelfarb, 2019). Education researchers use several test techniques from educational and psychological research in combination to analyse examination test data. Related research works and publications in education, psychology, and other sciences are continuously being published in journals. Several methods are used in test data analysis, including classical test theory, factor analysis, cluster analysis, item response theory, and model analysis (Ding & Beichner, 2009). These are summarized and presented in Table 2.

Table 2

Summary of Measurement Theories and Analytical Methods Applied in the Study

		Objectives	Implement
Classical test theory		Evaluate item or test reliability and discriminatory power	Perform item analysis and test analysis
Factor analysis	Principal component analysis	Reduce the number of variables	Solve eigenvalue equations for correlation matrix
	Common factor analysis	Explore underlying factors	Solve eigenvalue equations for adjusted correlation matrix
Cluster analysis		Classify subjects into groups	Calculate Euclidian distances and merge/divide subjects
Item response theory		Estimate item characteristics and subjects' latent abilities	Use logistic functions to formulate data
Model analysis		Represent probabilities of using different models	Calculate density matrix and solve eigenvalue equations

To determine whether the knowledge and skills of students who graduated from the PET program were reliably assessed, it was necessary to examine the difficulty levels and reliability of the tests and tasks included in the graduation examination. This was particularly important because the graduation examination covered content from all professional courses. To evaluate the task difficulty and test reliability, various theories and methods were considered (Table 2), with classical test theory being selected. Additionally, cluster analysis was employed to classify students according to their attitudes toward learning.

Classical Test Theory: Classical test theory is an important part of the foundation of modern measurement theory (Kline, 1986). The total test score consists of two components: the true test score and random error. Classical test theory provides the possibility to conduct several statistical analyses for test evaluation. This includes both item response and test analysis (Doran, 1980). The purpose of these analyses is to examine the reliability and discriminant index of any given test. In terms of test reliability, it considers whether the test produces the same results when administered twice (at different times), whether the test-taker's performance is consistent, and whether the testing conditions are similar.

Item analysis measures the following three aspects:

- Item difficulty level *(P)*
- Discriminant index *(D)*
- Point-biserial coefficient (r_{pbi})

In practice, item difficulty values between 0.3 and 0.9 are considered acceptable (Doran, 1980). If values fall above or below this range, it means the item is either too difficult or too easy. Discriminant index measures the differentiating ability between high-scoring and low-scoring students. In other words, it is the percentage difference in correct responses between students in the upper quartile and lower quartile for any given item (Oosterhof, 2001). Generally, discriminant index $D \geq 0.3$ is considered to meet standards (Doran, 1980). Higher values are better. If it is lower than this value, the item should be carefully reviewed and the question statement should be made clearer and more comprehensible. The point-biserial coefficient is a measure of the reliability of individual item, determined by the correlation between that item scores and total test scores (Ghiselli, 1981). A point-biserial coefficient $r_{pbi} \geq 0.2$ is considered acceptable (Kline, 1986). Similarly, higher values are better. If an item's coefficient is observed to be low, it indicates that the item does not have similar content to other items. Therefore, it is important to make improvements to that item by comparing it with other items.

Test Analysis: Test analysis can be determined through three measures: Kuder-Richardson reliability index (r_{test}), Ferguson's delta (δ), and Cronbach's alpha (a). These measures evaluate an entire test rather than assessing individual items. The Kuder-Richardson reliability index measures the internal consistency of a test. In other words, it evaluates whether the items in a test were developed within the same materials. When the correlation between items is high, the Kuder-Richardson reliability index is high. This indicates that the entire test has higher reliability. This is named and denoted as KR-20 after the equation number in Kuder and Richardson's famous article (Kuder & Richardson, 1937). If the r_{test} value is greater than 0.8, the test items are considered reliable. If the test reliability is low, items with low discriminant index and low point-biserial coefficients should first be examined. Because these items often are not consistent with other items, they can negatively impact the reliability of the entire test.

Ferguson's delta (δ) Ferguson's delta measures the discriminatory power of an entire test. Specifically, this measure studies how widely students' total scores are distributed across the possible range. Generally, the broader the score distribution is, the better the test is in discriminating among students at different levels. Ferguson's delta is calculated using the following formula. Generally, if Ferguson's delta is greater than 0.90, it is considered to effectively identify student differences (Kline, 1986). Cronbach's alpha is one of the comprehensive methods for examining the reliability of tests and research questionnaires. An alpha value greater than 0.5 is considered acceptable (Field, 2009).

Cluster Analysis: In 1971, Cormack, and in 1999, Gordon defined "cluster as internal cohesion (homogeneity) and external isolation (separation)" (Cormack, 1971; Gordon, 1999). Cluster analysis is a quantitative classification method. Cluster analysis is used to classify any objects into distinct groups based on their unique characteristics (Everitt et al., 2011). Cluster analysis often uses Euclidian distances to measure similarities between any

two subjects (Aldenderfer & Blashfield, 1984). Cluster analysis tends to address general problems in scientific fields such as Physics, Biology, Botany, Medicine, Psychology, Geography, Marketing, Image Processing, and Archaeology. There are many clustering methods and algorithms, among which the two-step clustering method from SPSS software was used. This method is based on a model that measures distances between neighbors in conditions involving both continuous and categorical variables. While most clustering methods are suitable for very large sample datasets, this method has the advantage of being applicable to small sample datasets.

To classify students' learning attitudes, a two-step cluster analysis was conducted using integrated assessments of knowledge and skills. Knowledge was represented by examination and test scores, while skills were represented by assignment-based performance scores obtained from compulsory courses, including general education, teacher education, professional, and practicum courses. These variables were selected to reflect students' sustained academic and practical engagement throughout the program.

The two-step cluster analysis method was chosen due to its suitability for handling continuous variables and its capacity to automatically determine the optimal number of clusters based on statistical criteria (Tkaczynski, 2017). In the first step, cases were pre-clustered using a distance-based approach, followed by hierarchical clustering to identify stable cluster solutions. This method allows for the identification of homogeneous groups of students with similar performance patterns in terms of knowledge and skills.

The resulting clusters were interpreted as distinct learning attitude groups, based on the assumption that consistent patterns of knowledge-skills performance across diverse learning contexts reflect students' orientations toward learning in practice-oriented disciplines such as PET.

The quality of the cluster solution was evaluated using the silhouette index, which measures the degree of similarity of cases within a cluster compared to those in other clusters. Silhouette values range from −1 to +1, with higher values indicating better cluster separation and cohesion. Values above 0.5 are generally interpreted as indicating good cluster quality, while values between 0.2 and 0.5 suggest reasonable separation (Davies & Bouldin, 1979; Shutaywi & Kachouie, 2021).

In this study, the silhouette index indicated an acceptable level of separation among the identified clusters, suggesting that the classification of students' learning attitudes based on knowledge and skills assessments was statistically meaningful. Visual inspection of cluster distributions further supported the distinctiveness of the clusters, as illustrated in Table 8.

Results

Results of Item and Test Analysis

The research results show that the graduation examination tests taken from students who graduated from the PET bachelor's program in the 2023-2024 academic year are overall appropriate for assessing students' knowledge and skills, with reliability indices exceeding acceptable values (Table 3). A total of 33 graduates took the graduation

examination in the winter semester and 34 in the spring semester. Out of the total 450 graduation examination test items, 50 randomly selected test items consisted of 41 closed-ended tests (multiple choice, sequencing, matching, closed-ended tests) and 9 open-ended written response tasks. The following table presents the detailed results of studying the difficulty level, discriminant power, and reliability of these 41 closed-ended test items.

Table 3

Item Difficulty, Discrimination Index, and Reliability Indicators of the selected graduation examination for PET program

		Graduation	
		Winter	**Spring**
Test statistics	Critical values	Calculated values	
Difficulty level index, *P*	[0.3, 0.90]	0.79*	0.81*
Discrimination index, *D*	≥0.30	0.54*	0.57*
The point biserial coefficient, r_{pbi}	≥0.20	0.53*	0.55*
Kuder-Richardson reliability index (KR-20)	≥0.70	0.96	0.93
Reliability index (Cronbach's α)	≥0.5	0.88	0.93
Ferguson's delta, δ	≥0.90	0.97	0.98

Note: *P* = difficulty index; D = discrimination index; r_{pbi} = point-biserial coefficient; KR-20 = Kuder-Richardson reliability index; α = Cronbach's alpha; δ = Ferguson's delta. *Values indicated by average across items.

Because the difficulty levels index and reliability of the 41 randomly selected test items included in the graduation examination were found to be acceptable, the grades assessing the knowledge and skills of students enrolled in the PET program were considered sufficiently reliable to support subsequent clustering analysis. On this basis, further analyses were conducted with the aim of classifying students according to their learning attitudes.

Cluster Analysis Results

Students who enrolled in the PET program in 2020 and graduated in 2024 were classified by learning attitudes based on their knowledge and skills assessments in compulsory courses including general, teacher education, professional, and practicum courses. No survey was administered to measure the learning attitudes of the 2024 graduates, the study sought to infer students' attitudes from their demonstrated scores of knowledge and skills, as also described in the research data section this article. A two-dimensional space was constructed, with knowledge scores represented on the horizontal axis and skills scores on the vertical axis. Each cluster formed within this space represents a distinct level or group of attitudes toward learning, such that graduates within the same cluster exhibit similar attitude. In summary, learning attitude was operationalized as the integrated pattern of students' knowledge and skills performance.

The PET program consists of four main sections: general, teacher education, professional, and practicum courses. General courses included 9 courses totaling 19 credits, such as Mongolian History and Culture, Mongolian Language and Stylistics, and Information and Communication Technology. Teacher education courses included 6 courses totaling 14 credits, such as General Psychology, and Foundations of Learning and

Teaching. Professional courses included 34 courses totaling 73 credits, such as Theory of Physical Education, Sports Medicine, and Sports Skills. Practicum courses included 13 credits of Orientation Practicum, Study Practicum, Guided Practicum, and Teaching Practicum.

Analysis was conducted using the two-step clustering method on the knowledge and skills results from a total of 53 courses. At MNUE, the final grade for each course is calculated as the aggregate of three assessment components: attendance, class participation, examinations (knowledge), and assignments (skills). First, analysis was performed on general courses, followed by teacher education, professional, and practicum courses. Subsequently, the average percentage of knowledge and skills from all compulsory courses was calculated and used to classify bachelor's students' learning attitudes into 5 groups. The results classified by the two-step clustering algorithm are summarized and presented in detail in Table 4.

Table 4

Two-Step Cluster Analysis Results of Learning Attitudes Based on Knowledge and Skills Performance Across Compulsory Course Categories. The blue curve is indicated by 5 groups distributions, and the blue curve represents a group's distribution.

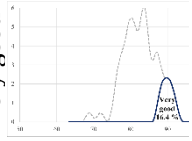
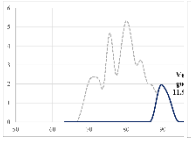
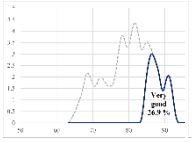
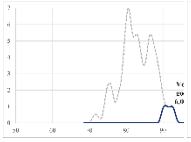
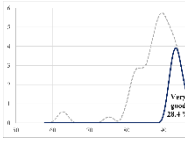
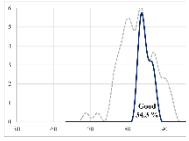
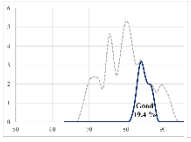
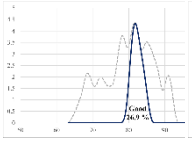
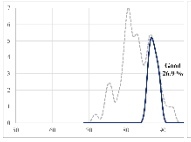
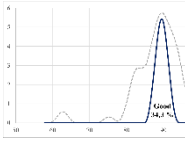
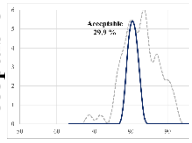
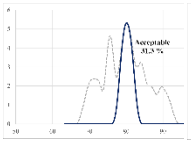
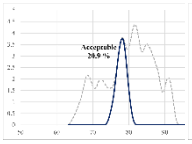
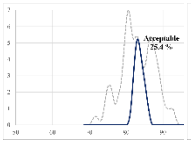
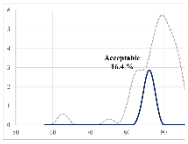
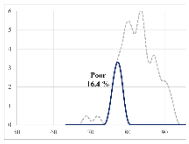
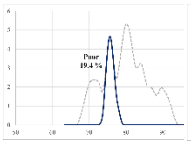
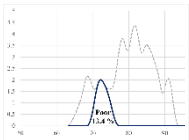
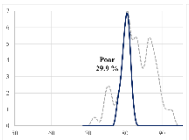
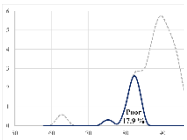
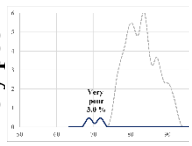
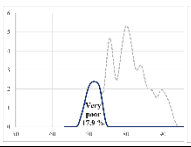
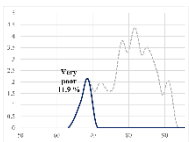
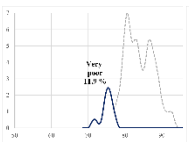
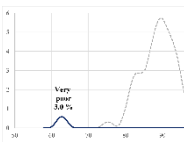
Groups	All courses	General courses	Teacher Education courses	Physical Education courses	Practices
Very good					
Good					
Acceptable					
Poor					
Very poor					

Table 5 shows the classification of students' learning attitudes into 5 groups - very good, good, average, poor, and very poor - using cluster analysis method based on the average scores of knowledge and skills from General, Teacher Education, Professional, Practicum and All Courses, according to whether students' learning attitudes were well developed or not. The percentage and graph of each group are shown respectively. When classifying students by the average percentage of all courses, they were distributed as follows: very good 16.4%, good 34.3%, average 29.9%, poor 16.4%, and very poor 3.0% in their respective groups.

Table 5

Means and Weights for 5 groups across Compulsory Course Categories

Groups	All Courses		General Courses		Teacher Education Courses		Professional Courses		Practices	
	Mean	%	Mean	%	Mean	%	Mean	%	Mean	%
Very good	89.8	16.4	90.2	11.9	88.2	26.9	91.5	6.0	93.9	28.4
Good	84.8	34.3	84.8	19.4	82.3	26.9	87.4	26.9	89.5	34.3
Acceptable	80.6	29.9	80.21	31.3	77.9	20.9	83.6	25.4	86.1	16.4
Poor	77.2	16.4	75.8	19.4	72.7	13.4	80.2	29.9	81.7	17.9
Very poor	70.3	3.0	71.1	17.9	67.9	11.9	75.1	11.9	62.8	3.0

Note: Mean scores represent the average percentage of knowledge and skills assessments. % = percentage of graduates classified in each group. Groups are ordered from highest (Very good) to lowest (Very poor) learning attitude level.

When examining the arithmetic mean of the distribution of students' learning attitudes classified by cluster analysis, the following numerical values were obtained. Table 6 shows the mean values and standard deviations of students' knowledge and skills assessments across General, Teacher Education, Professional, Practicum, and All Courses. Around and above the overall course average score (82.69), students with average (29.9%), good (34.3%), and very good (16.4%) learning attitudes were developed. The sum of these three groups, or 80.6% of students, can be concluded to have developed learning attitudes. The averages for practicum courses were relatively high. This is understood to be related to students' interest in teaching activities.

Table 6

Mean and standard deviation grades of students for Compulsory Course Categories

	N	Mean	Std. Deviation
All Courses	67	82.69	4.77
General Courses	67	79.81	6.06
Teacher Education Courses	67	79.93	6.87
Professional Courses	67	83.05	4.57
Practicum	67	87.96	6.33

When performing two-step cluster analysis on SPSS software, it calculates coefficient values that express how significant the clustering is. Table 7 shows the importance of clustering performed by courses and the average of all courses. Of course, the closer the significance is to 1, the better it indicates that the groups were well differentiated.

Table 7

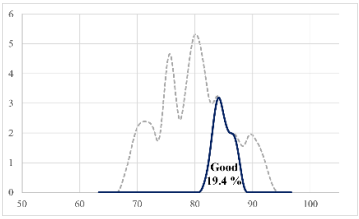
Silhouette index and input (predictor) importance of Two steps cluster component

	All Courses	General Courses	Teacher Education Courses	Professional Courses	Practices
Silhouette measure of cohesion and separation	0.79*	0.71*	0.7*	0.8*	0.69*

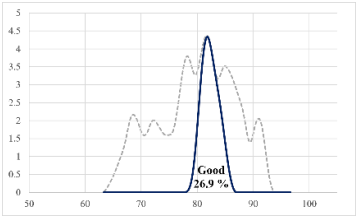
Note: Silhouette values range from −1 to +1; values above 0.5 indicate good cluster separation and cohesion. *Input (predictor) importance = 1.0 for all course categories.

When observing the distribution patterns of average assessments for General, Teacher Education, and Practicum courses, the groups do not appear to be clearly differentiated, whereas the 5 groups are very clearly differentiated when looking at Professional courses (silhouette = 0.8) and all courses (silhouette = 0.79). This can be observed in the gray graphs presented in Table 7.

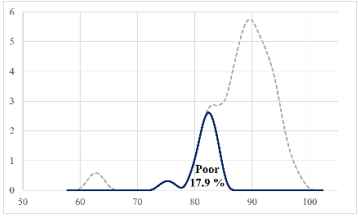
Figure 2 shows the distributions of students having a good level of general academic achievements have positive learning attitudes. It was shown separately in General Courses, Teacher Education Courses and Practicum. The grey curves indicate 5 group distributions, and the blue curve represents the distribution for students having a good level of general academic achievements have positive learning attitudes. The horizontal axis represents students' mean knowledge and skills scores.



General Courses



Teacher Education Courses



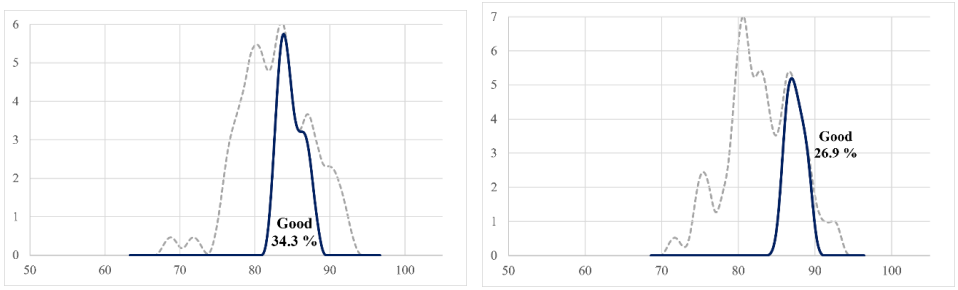
Practices

Figure 2: Distributions of students' levels of general academic achievements (1)

To make it more comprehensible, let's examine the graphs of students with well-developed learning attitudes one by one. The graphs created by performing cluster analysis on the scores of General, Teacher Education, and Practicum courses are presented in Figure 2. Observing these 3 graphs, it can be seen that performing cluster analysis with relatively few courses and classifying students by their learning attitude development is unclear, meaning it cannot differentiate well.

The results obtained by finding the average of Professional and All courses for each student and performing cluster analysis, expressed through distribution graphs for each of the 5 groups - very good, good, average, poor, and very poor are shown in Figure 3. This demonstrates from the grey-colour graph in Figure 3 that when classifying students by learning attitudes, it is appropriate to consider them as a sum and complex of quite a few courses. This is because our 5 groups - very good, good, average, poor, and very poor - are clearly differentiated on the graphs of 31 Professional courses and all 49 courses. This is expressed by the cluster analysis significance shown in Table 8, which is 0.8 for Professional courses and 0.79 for All courses.

Figure 3 shows the distributions of students having general academic achievements have learning attitudes. Their learning attitudes were shown separately in 5 groups: very good, good, acceptable, poor and very poor. The grey curve is indicated by 5 group distributions, and the blue curve represents the distribution for the group students. The horizontal axis represents the average values of students knowledge and skills.



All Courses **Professional Courses**

Figure 3: Distributions of students' levels of general academic achievements (2).

Table 8 presents the distribution of learning attitude levels among the 67 students who successfully completed the PET program during the 2023–2024 academic year. The findings indicate that 11 students (16.4%) demonstrated very good learning attitudes, 23 students (34.4%) exhibited good learning attitudes, and 20 students (29.9%) showed average learning attitudes. In contrast, 11 students (16.4%) were categorized as having poor learning attitudes, while 2 students (3.0%) demonstrated very poor learning attitudes.

Table 8

Distribution of PET Program Graduates Across Five Learning Attitudes Groups (2023–2024 Academic Year)

	Very good	Good	Acceptable	Poor	Very poor
All Courses	11	23	20	11	2
	16.4 %	34.4 %	29.9 %	16.4 %	2.9 %

In the 2023-2024 academic year, 80.7% of graduates from the PET program had higher learning attitudes, while 19.3% had lower learning attitudes. It is considered necessary to include graduates with insufficiently developed learning attitudes in professional development training in the future.

Discussion

The findings of this study are consistent with established standards in Classical Test Theory and align with comparable program evaluation studies in Physical education. The graduation examination items demonstrated acceptable difficulty indices (P = 0.79–0.81), discrimination indices (D = 0.54–0.57), and point-biserial coefficients (r_{pbi} = 0.93–0.96), as well as high Kuder-Richardson reliability (KR-20 = 0.93–0.96) and Ferguson's delta (δ = 0.97–0.98). These values exceed the minimum thresholds recommended by Doran (1980) and Kline (1986), indicating that the assessments reliably differentiated student performance levels. In comparison, Sankaran and Saad (2022) employed the CIPP model to evaluate a Bachelor of Education program in Malaysian polytechnics and similarly found that while institutional assessments were structurally sound, limitations in instructional resources constrained the full realization of program outcomes. Iqbal et al. (2022) likewise applied the CIPP framework to evaluate an undergraduate Health and Physical Education program and identified gaps in the process component, including mentorship and physical activity arrangements, rather than in assessment reliability per selected items. The present study advances this line of inquiry by demonstrating that, within the Product component of the CIPPIEST model, graduation examination data can serve as a credible and psychometrically defensible basis for outcome evaluation, even in the absence of externally standardized instruments.

The classification of students' learning attitudes into five distinct clusters—very good, good, acceptable, poor, and very poor based on integrated knowledge and skills performance represents a methodologically distinctive contribution to program evaluation in PET. The silhouette indices for professional courses (0.80) and all courses combined (0.79) confirmed acceptable cluster cohesion and separation, indicating that the two-step cluster analysis effectively differentiated student profiles. This approach responds to a recognized limitation in the field: as noted by Halim et al. (2022) and Ernest (2015), learning attitudes, knowledge, and skills are closely interrelated constructs, yet most existing studies measure them separately using self-report instruments. The present study's use of longitudinal academic performance data as a proxy for learning attitudes offers an alternative operationalization that is both feasible and empirically grounded. Chu et al. (2022) similarly found that pedagogical models in Physical education significantly influence students' affective engagement and learning motivation, reinforcing the premise that academic performance patterns reflect deeper attitudinal orientations. Furthermore, Şenay (2013) demonstrated that the integrated influence of learning attitudes, knowledge, and digital skills substantially determines University students' engagement in modern educational environments, a finding that supports the theoretical basis of the integrated performance-learning attitudes framework adopted in this study. Compared to survey-based studies, the cluster-based approach applied here provides a more objective and reproducible classification method, particularly suited to institutional contexts where retrospective learning attitudes measurement is not feasible.

From a program evaluation standpoint, the distribution of graduates across learning attitude clusters 80.7% classified within higher-attitude groups and 19.3% in lower-attitude groups carries direct implications for curriculum reform and targeted professional development in the PET program. Maulana (2024) demonstrated that CIPP-based evaluation of Physical education programs in Indonesian high schools facilitated

systematic and continuous improvements in student learning, while Mulyadi et al. (2024) emphasized that effective curriculum reform in Physical education requires differentiated, context-sensitive strategies supported by sustainable teacher development initiatives.

The present findings resonate with these conclusions: the identification of graduates with comparatively lower learning attitudes highlights the need for individualized follow-up interventions, including mentorship, remedial coursework, and enhanced practicum supervision. Notably, the relatively high performances scores observed in practicum courses (mean = 87.96) compared to General and Teacher education courses suggest that students demonstrate stronger engagement in applied, field-based learning contexts, which is consistent with the nature of practice-oriented disciplines.

Within the CIPPIEST framework, these results provide evidence-informed foundations for revising Program Learning Outcomes, realigning curricular priorities, and designing targeted support structures particularly as the PET program at MNUE undergoes comprehensive curriculum reform. Taken together, the study demonstrates that performance-based cluster analysis constitutes a valid and institutionally practical method for evaluating learning attitudes as a component of program outcomes, offering a complementary approach to conventional survey-based evaluations in Teacher Education teacher program.

Conclusion

This study extends the application of the CIPPIEST evaluation framework by demonstrating how program outcomes in Physical Education Teacher (PET) can be systematically examined using validated assessment data and cluster-based classification methods. The findings indicate that the CIPPIEST model provides a structured basis for evaluating knowledge, skills, and learning attitudes as integral components of program outcomes.

The psychometric analysis of randomly selected graduation examination items confirmed that the tests met acceptable standards of difficulty (p = 0.79–0.81), discrimination (D = 0.54–0.57), and reliability (r_{pbi} = 0.93–0.96), supporting the validity of the performance data used in subsequent analyses. Building on these validated measures, the study demonstrated that learning attitudes related to academic and professional education may be meaningfully inferred from integrated patterns of knowledge and skills performance across multiple courses.

The two-step cluster analysis identified five distinct learning attitudes among graduates. The overall course mean score of 82.69 reflects a generally strong level of cumulative knowledge and skills achievement within the program. Furthermore, 80.7% of graduates were classified within higher learning attitude clusters, while 19.3% were grouped into comparatively lower profiles. This distribution underscores the potential value of designing targeted professional development and follow-up support mechanisms for graduates identified in lower learning attitude clusters, thereby contributing to continuous program improvement.

Nevertheless, this study is subject to certain limitations, including its focus on a single institution and the reliance on graduation examination data, which may not fully capture

the breadth of graduate competencies; future research should therefore expand the scope to include multiple teacher education programs and incorporate longitudinal follow-up data to examine the predictive validity of learning attitudes in professional practice. The findings carry practical implications for program designers and academic administrators, suggesting that embedding cluster analysis based performance analytics within routine program Evaluation cycles can enable more responsive, evidence-informed curriculum development and graduate support strategies in bachelor program.

Acknowledgments

We express our deep gratitude to the Research and Innovation Department of MNUE for funding the commissioned project on program evaluation (the project number MNUE2025C004). We also specially thank IT engineers Ts.Buyanjargal and M.Munkhsaikhan for extracting student data from ULMS for the PET program.

References

Aitana, G.-O.-d.-Z., Alonso-García, M. A., Gómez-Flechoso, M. d. l. Á., & Aliagas, I. (2025). Peer Mentoring, University Dropout and Academic Performance before, during, and after the Pandemic in Spain. *Evaluation and Program Planning*, *113*, 102676. https://doi.org/10.1016/j.evalprogplan.2025.102676

Aldenderfer, M., & Blashfield, R. (1984). Cluster Analysis. In: SAGE Publications, Inc. http://dx.doi.org/10.4135/9781412983648

Alquraan, M. F., Alazzam, S., & Farhat, D. (2025). Assessing the Structural Relationships between the Cipp Model Components in Teacher Education Program. *Asian Education and Development Studies*, *14*(1), 103-114. https://doi.org/10.1108/aeds-09-2024-0217

Bellido-Medina, R., Lazo-Manrique, M. C., Lazo-Manrique, A. P., Asillo-Apaza, Y. Y., Martinez-Martinez, R. L., Aguilar-Del Carpio, D. E., Bedoya-Zaira, G. A., Banda-Cardenas, J. D., & Calizaya-Lopez, J. (2023). Attitude, Motivation, Anxiety, and Academic Performance during the Learning Process in Students at Public Universities in Peru. *Journal of Higher Education Theory and Practice*, *23*(15). https://doi.org/10.33423/jhetp.v23i15.6430

Chu, Y., Chen, C., Wang, G., & Su, F. (2022). The Effect of Education Model in Physical Education on Student Learning Behavior. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.944507

Cormack, R. M. (1971). A Review of Classification. *Journal of the Royal Statistical Society. Series A (General)*, *134*(3), 321. https://doi.org/10.2307/2344237

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224-227. https://doi.org/10.1109/tpami.1979.4766909

Ding, L., & Beichner, R. (2009). Approaches to Data Analysis of Multiple-Choice Questions. *Physical Review Special Topics - Physics Education Research*, *5*(2). https://doi.org/10.1103/physrevstper.5.020103

Doran, R. L. (1980). *Basic Measurement and Evaluation of Science Instruction*. ERIC. https://lccn.loc.gov/81119723

Erjing, N. (2025). A Comparative Study on the Development of Vocational Education in Mongolia and China: Pathways, Challenges, and Synergies. *Journal of Exploration of Vocational Education*, *7*(6), 33-67. https://doi.org/10.63650/jeve.v7i6.96

Ernest, A. (2015). Relationship of Students' Attitudes toward Science and Academic Achievement. In *Attitude Measurements in Science Education* (pp. 245-262): Emerald Publishing Limited. http://dx.doi.org/10.1108/978-1-68123-086-320251012

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis*. Wiley. https://doi.org/10.1002/9780470977811

Field, A. (2009). *Discovering Statistics Using Spss* (3rd ed.). SAGE Publications. https://books.google.com.pk/books?id=a6FLF1YOqtsC

Ganbaatar, U., Gantulga, O., Byambajav, P., Och, M., Ganburged, G., Jadamba, T., Dagvajantsan, B., & Byambasukh, O. (2021). Relationship of Tooth Loss to Mild Cognitive Impairment among Middle-Aged Mongolians: Mon-Timeline Study. *Neuroscience Research Notes*, *4*(4), 10-18. https://doi.org/10.31117/neuroscirn.v4i4.88

Ghiselli, E. E. (1981). Measurement Theory for the Behavioral Sciences. In J. P. Campbell & S. Zedeck (Eds.). San Francisco :: W. H. Freeman. https://searchworks.stanford.edu/view/1482546

Gordon, A. D. (1999). *Classification* (2nd ed.). CRC Press. https://books.google.com.pk/books?id=_w5AJtbfEz4C

Halim, A., Wirda, A., & Yusrizal, Y. (2022). Analysis of Learning Styles in Terms of Knowledge, Skills and Attitudes. *Momentum: Physics Education Journal*, *6*(2), 162-170. https://doi.org/10.21067/mpej.v6i2.6581

Himelfarb, I. (2019). A Primer on Standardized Testing: History, Measurement, Classical Test Theory, Item Response Theory, and Equating. *Journal of Chiropractic Education*, *33*(2), 151-163. https://doi.org/10.7899/jce-18-22

Iqbal, Z., Khan, R., & Wadood, A. (2022). Evaluation of the Effectiveness of the Process of Undergraduate Health and Physical Education Program by the Cipp Model. *Global Educational Studies Review*, *VII*(II), 285-295. https://doi.org/10.31703/gesr.2022(vii-ii).27

Kline, P. (1986). *A Handbook of Test Construction : Introduction to Psychometric Design*. London ; New York, NY : Methuen. https://archive.org/details/handbookoftestco0000klin

Kuder, G. F., & Richardson, M. W. (1937). The Theory of the Estimation of Test Reliability. *Psychometrika*, *2*(3), 151-160. https://doi.org/10.1007/bf02288391

Maulana, S. (2024). Evaluation of the Cipp Model in the Implementation of Ktsp Physical Education Learning. *Sport Pedagogy Journal*, *13*(1), 32-41. https://doi.org/10.24815/spj.v13i1.38493

Mulyadi, M., Hidayatullah, M. F., Syaifullah, R., & Riyadi, S. (2024). Evaluative Study of the Implementation of the Independent Curriculum in Physical Education Elementary School Level Using the Cipp Model. *QALAMUNA: Jurnal Pendidikan, Sosial, dan Agama*, *16*(2), 1017-1030. https://doi.org/10.37680/qalamuna.v16i2.5718

Musa, S., Mamudo, W. S., Mohammed, B. B., & Audu, J. H. (2022). Correlation between Students' Attitudes and Mathematics Learning Achievements of High School Students in Yobe, Nigeria. *Indonesian Journal of Science and Mathematics Education*, *5*(2), 147-155. https://doi.org/10.24042/ijsme.v5i2.11273

Oosterhof, A. (2001). *Classroom Applications of Educational Measurement* (3rd ed.). Upper Saddle River: Merrill. https://cir.nii.ac.jp/crid/1971993809682497965

Sankaran, S., & Saad, N. (2022). Evaluating the Bachelor of Education Program Based on the Context, Input, Process, and Product Model. *Frontiers in Education*, *7*. https://doi.org/10.3389/feduc.2022.924374

Şenay, Ş. H. (2013). The Attitudes of University Students Towards Learning. *Procedia - Social and Behavioral Sciences*, *83*, 947-953. https://doi.org/10.1016/j.sbspro.2013.06.177

Setyadi, F., Hidayatullah, M. F., & Purnama, S. K. (2022). Evaluation of Sports and Health Physical Education Program in Sma N 2 Ngawi Using the Cipp Evaluation Model (Context, Input, Process, Product). *Indonesia Sport Journal*, *4*(2), 36-42. https://doi.org/10.24114/isj.v4i2.37872

Shutaywi, M., & Kachouie, N. N. (2021). Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. *Entropy*, *23*(6), 759. https://doi.org/10.3390/e23060759

Stufflebeam, D. L. (2003). The Cipp Model for Evaluation. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International Handbook of Educational Evaluation* (pp. 31-62). Springer Netherlands. https://doi.org/10.1007/978-94-010-0309-4_4

Stufflebeam, D. L. (2015). *Cipp Evaluation Model Checklist: A Tool for Applying the Cipp Model to Assess Projects and Programs* (2nd ed.). Western Michigan University. http://rszarf.ips.uw.edu.pl/ewalps/dzienne/cipp-model-stufflebeam2015.pdf

Suryobroto, A. S., Ani Hastuti, T., & Maya Jatmika, H. (2018). *Using the Context, Input, Process, and Product Evaluation Model (Cipp) to Evaluate Elementary School Teacher-Learner Program of Physical Education in Yogyakarta City* Proceedings of the 2nd Yogyakarta International Seminar on Health, Physical Education, and Sport Science (YISHPESS 2018) and 1st Conference on Interdisciplinary Approach in Sports (CoIS 2018), http://dx.doi.org/10.2991/yishpess-cois-18.2018.58

Tkaczynski, A. (2017). Segmentation Using Two-Step Cluster Analysis. In T. Dietrich, S. Rundle-Thiele, & K. Kubacki (Eds.), *Segmentation in Social Marketing: Process, Methods and Application* (pp. 109-125). Springer Singapore. https://doi.org/10.1007/978-981-10-1835-0_8

Toosi, M., Modarres, M., Amini, M., & Geranmayeh, M. (2021). Context, Input, Process, and Product Evaluation Model in Medical Education: A Systematic Review. *Journal of Education and Health Promotion*, *10*(1). https://doi.org/10.4103/jehp.jehp_1115_20

Tudevdagva, U., & Sanjdorj, T. (2022). The Evaluation Result of Online Master Course during Pandemic: Case of Industrial Management Course at Mongolian University of Science and Technology. *Embedded Selforganising Systems*, *9*(3), 31-36. https://doi.org/10.14464/ess.v9i3.536

Umac, N. L., Panara-Ag, D. S., & Yurango, C. P. (2025). The Relationship between Students' Attitudes Towards Mathematics and Academic Performance in the Subject. *Asian Journal of Education and Social Studies*, *51*(8), 709-719. https://doi.org/10.9734/ajess/2025/v51i82271

Wahidah, I., Listyasari, E., Rahmat, A. A., & Rohyana, A. (2023). Evaluation of Physical Education Independent Curriculum through Cipp: Managerial Implementation in Learning Activities. *Indonesian Journal of Sport Management*, *3*(2), 208-223. https://doi.org/10.31949/ijsm.v3i2.6403