International Journal of Social Sciences and Management Review

Volume: 09, Issue: 02 March - April 2026

ISSN 2582-0176

VALIDITY AND RELIABILITY OF PHYSICAL EDUCATION TEACHING IMPLEMENTATION INSTRUMENTS IN SECONDARY SCHOOLS

SYED KAMARUZAMAN SYED ALI[1], ANNISAA BASAR[1], NOVRI GAZALI[2] & M. FRANSAZELI MAKOROHIM[2]

University of Malaya[1], Riau Islamic University[2]

https://doi.org/10.37602/IJSSMR.2025.9208

ABSTRACT

This study aims to evaluate the validity and reliability of the physical education (PE) teaching implementation instrument in secondary schools. A valid and reliable instrument is a fundamental basis for ensuring data accuracy and teaching effectiveness in fostering an active lifestyle among students. This study employs a quantitative research design in the form of a survey involving physical education teachers. The instrument used is a structured observation checklist covering four main constructs: implementation of teaching strategies, use of resource materials, use of facilities, and implementation of assessment. The content validity process involved two phases: a review by six experts in the field of Physical Education and a second-phase evaluation by PE teachers. The reliability of the instrument was analyzed using the Kuder-Richardson 21 (KR-21) formula and inter-rater reliability. Results from the pilot study showed high reliability coefficients for all dimensions, with values ranging from 0.600 to 0.985, while the inter-rater reliability value for teaching strategies reached 91.7%. The research findings conclude that this instrument has a satisfactory level of validity and reliability for measuring the PE teaching process systematically. The implications of the study suggest this instrument as a standardized evaluation tool to help schools identify the strengths and weaknesses of teacher instruction while supporting continuous professional development.

Keywords: Validity, Reliability, Physical Education, Structured Observation, Teaching Instrument.

1.0 INTRODUCTION

Physical Education in secondary schools plays a vital role in fostering an active lifestyle, improving physical health, and supporting the social and emotional development of students. To ensure the effectiveness of instructional delivery, teachers must utilize teaching implementation instruments that are both valid and reliable. The validity and reliability of an instrument not only guarantee the accuracy of the collected data but also ensure that the evaluation of teaching implementation truly reflects the reality of the classroom. Therefore, research into these aspects is crucial for strengthening the quality of physical education in secondary schools.

In the context of educational research, valid and reliable instruments serve as the foundation for authoritative findings. Without validity, an instrument may fail to measure the intended construct, while without reliability, measurement results may fluctuate and remain

Volume: 09, Issue: 02 March - April 2026

ISSN 2582-0176

inconsistent. Thus, this study aims to evaluate the validity and reliability of physical education teaching implementation instruments in secondary schools, so they may serve as a solid reference for enhancing teaching effectiveness and supporting more systematic and efficient curriculum development. Furthermore, the research objectives are as follows:

1.1 Research Objectives:

1. To evaluate the validity of teaching implementation instruments for secondary school Physical Education.
2. To evaluate the reliability of teaching implementation instruments for secondary school Physical Education.

1.2 Related Studies

Several studies have been conducted to evaluate research instruments. D'Aoust R. (2025) developed a course evaluation instrument in nursing education, focusing on validation to align teaching effectiveness with student learning outcomes. The methodology involved item construction based on scientific criteria and reliability testing through factor analysis. The instrument was designed to measure aspects such as clear goals, adequate preparation, and appropriate teaching methods. Findings indicated that the instrument is valid and reliable, showing high alignment between teaching evaluation and student outcomes, such as improvements in clinical skills. This supports the use of the instrument for continuous improvement. Educational institutions can utilize similar instruments for formative assessment to ensure effective teaching and optimal learning outcomes.

Meanwhile, Medina M.S. et al. (2019) examined strategies for designing teaching evaluation instruments based on student ratings, consistent with scholarly teaching criteria such as clear goals and appropriate methods. The recommended instrument contains 10-20 rating scale questions and one written response question, emphasizing question clarity and flexibility for specific courses. Administration involves the use of online tools for efficiency, random student selection, and end-of-semester evaluations to increase response rates. Results showed that evaluations are effective when implemented voluntarily and without coercion. The use of these results requires a clear assessment plan, including faculty reflection and action plans for improvement, ensuring that evaluations support continuous professional development rather than punitive decisions.

Furthermore, Post, J. (2014) conducted a study on teacher evaluation instruments and processes, focusing on strengths and weaknesses. The methodology involved a literature review and examples of instruments, such as observation forms. Findings showed that the instruments effectively measure teaching skills and development but require validation for reliability. The study suggested that teachers use these instruments for self-assessment and quality instruction.

Lastly, Charles A. Peck, Maia Goodman Young, & Wenqi Zhang (2021) conducted a study on the use of teaching performance assessments for the evaluation and improvement of teacher education programs through authentic artifacts such as lesson plans and videos. The methodology involved historical reviews and implementation studies. Findings indicated that theTeaching Performance Assessment supports teacher candidate learning and program

Volume: 09, Issue: 02 March - April 2026

ISSN 2582-0176

improvement when used formatively, though it depends on organizational support and state policies.

Figure 1: Conceptual Framework

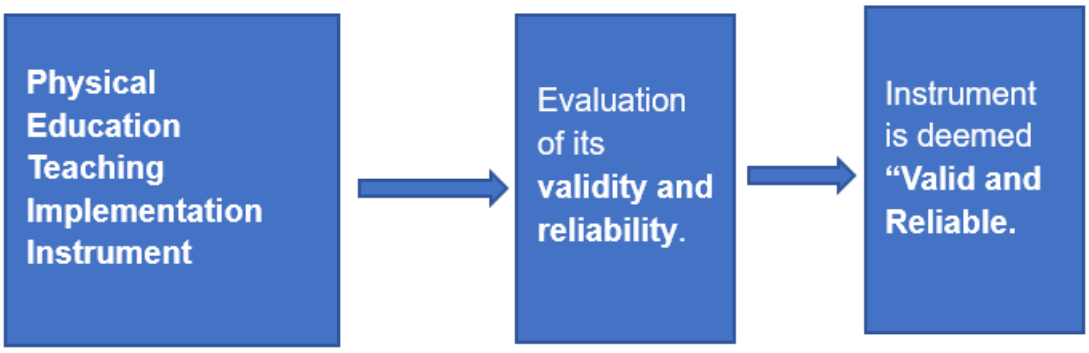

Figure 1, illustrates a clear process flow for evaluating and validating the Physical Education teaching implementation instrument. Referring to Figure 1, the "Physical Education Teaching Implementation Instrument" is a tool developed to assess the effectiveness of a teacher's instruction in Physical Education. Once developed, it undergoes an evaluation of its validity and reliability.

The purpose is to ensure that the instrument accurately measures the intended aspects of Physical Education teaching (validity) and produces consistent results (reliability). Ultimately, the instrument is deemed "Valid and Reliable." This outcome represents the result of the evaluation process, indicating that the instrument has been verified and can be used with confidence in an educational context.

2.0 METHODOLOGY

This study employs a quantitative research design in the form of a survey. The primary focus is to evaluate the validity and reliability of the teaching implementation instrument for physical education in secondary schools.

The research population consists of physical education teachers in secondary schools. Samples were selected using a stratified random sampling method to ensure representation based on school zones, teaching experience, and academic backgrounds. The sample size was determined based on the Krejcie & Morgan (1970) formula.

2.1 Research Instruments

An instrument is any tool used to obtain information, measure, or study a particular matter. This research utilizes a structured observation instrument developed based on the main constructs of "Implementation of Teaching and Learning Strategies," "Use of Teaching Materials and Resources," "Use of Facilities and Equipment," and "Implementation of Evaluation/Assessment". The instrument was constructed based on related theories, relevant studies, the Physical Education syllabus and its descriptions, Physical Education textbooks,

Volume: 09, Issue: 02 March - April 2026

ISSN 2582-0176

information from Physical Education committee files, as well as interviews and document analysis conducted by the researcher at several secondary schools in Selangor.

Structured observation can utilize checklists, rating scales, and rubrics. According to Gronlund (2003), observation is a general method that can be used to view achievement results, while checklists and rating scales are used to assist in conducting an observation to identify whether each action has been implemented or not. Observation using a checklist can reduce bias and errors. Observation via checklists has previously been used by Noraini, Loh, Ahmad Zabidi, Norjoharuddeen & Rahimi (2006) in research related to the implementation of Mathematics and Science teaching and learning in secondary schools.

Rating scales can also be used to determine the degree of use of a particular element. Rating scales are based on the frequency of a performed action, for example, "always," "sometimes," and "never". Checklists and rating scales are also lists capable of measuring dimensions related to outcomes. Accordingly, in this study, observation in the form of checklists and rating scales is used to obtain information related to the implementation of Physical Education (physical fitness) teaching and learning. The structured checklist instrument for this observation method was built based on related theories, relevant studies, Form 4 Physical Education textbooks, the Physical Education subject syllabus, Form 4 Physical Education syllabus descriptions, and information from secondary school Physical Education committee files.

The observation method used to collect information in this study is the non-participant observation method. The non-participant observation method only requires the researcher to remain outside the social unit whose behavior is being observed. The researcher functions as an observer without interfering with the behavior of the social unit. The researcher acts as an observer, listener, and recorder of behavior. Information collected through this method is more reliable because it is less influenced by the subject's reactions and less influenced by the researcher's feelings and sympathy.

3.0 RESEARCH FINDING

3.1 Validity and Reliability of the Instruments

The validity and reliability of an instrument are essential procedures in the development of any research instrument. Therefore, before the instrument was used for the actual study, the researcher conducted validity and reliability tests on the instrument beforehand.

3.2 Validity of the Instruments

Validity is a fundamental element in quantitative research. It is a crucial concept in the context of measuring constructs or ideas such as attitude, motivation, perception, fitness, and achievement. Validity can be divided into two types: external validity and internal validity. According to Balachandher (2005), internal validity encompasses content validity, criterion-related validity, and construct validity. To obtain validity for the observation instrument (structured checklist), the researcher conducted content validity.

In establishing content validity, the researcher, under the guidance of a supervisor, reviewed and carefully read the sentences, words, or sentence structures to determine whether they were

Volume: 09, Issue: 02 March - April 2026

ISSN 2582-0176

appropriate for the measurement components. This validity is necessary to ensure that the instrument used actually measures the concepts intended to be measured. For example, when the study is conducted to observe the implementation process of Physical Education (physical fitness) teaching, the researcher includes all topics typically covered in the Physical Education (physical fitness) curriculum. This is done so that the instrument can truly measure what is required.

In measuring the construct or implementation of Physical Education (physical fitness) teaching, the implementation process dimension is given attention when selecting items. Therefore, content validity is conducted to examine whether the question items or statements are suitable for measuring the implementation of Physical Education (physical fitness) teaching. These items are examined in terms of their phrasing, focus, and the terminology used. The more items that represent the dimensions of a construct or concept, the better the content validity.

After researching the questions or statements in the instrument, the researcher implemented a two-phase process to obtain content validity, similar to the study conducted by Gurvitch, Blankenship, Metzler, & Lund (2008). In the first phase, the researcher sent the instrument to six panel members who possess expertise in the fields of Physical Education, measurement, and program evaluation so that the panel could comment on the statements found in the instrument and subsequently provide feedback regarding its wording and content. This is because developing an instrument is highly complex and is a technical task that requires the assistance of experts in related fields.

Azizi, Shahrin, Jamaludin, Yusof & Abdul Rahim (2007) also emphasized that determining content validity is a judgment that can be made by using panel members to weigh the extent to which the instrument meets the standards. According to Thomas & Nelson (2001), content validity does not require statistical evidence but is sufficient by obtaining the views and opinions of experts in the field. After receiving the instruments reviewed by the panels, the researcher re-examined the instrument and made corrections based on the panel's criticisms and suggestions. For the second phase of this validity process, the researcher distributed the improved instrument to two Physical Education teachers to be completed. While providing their responses, the Physical Education teachers were also encouraged to make any notes regarding errors or statements that remained unclear concerning the content of the instrument. The second phase serves as the final check in the validity process. The respondents involved in this second phase of the validity process were not included as respondents in the actual study.

3.3 Reliability of the Instruments

Reliability is a fundamental element in quantitative research. In this study, instrument reliability is measured using Cronbach's Alpha and Inter-rater Reliability analysis. According to Sekaran (2000), Cronbach's Alpha is a coefficient or reliability value that indicates how the research items correlate with one another. A coefficient or index value approaching 1.0 signifies high reliability. Meanwhile, the minimum acceptable reliability coefficient or index is 0.60. The Cronbach's Alpha analysis process was conducted using a computer via the Statistical Package for the Social Sciences (SPSS Version 12.0) program.

Volume: 09, Issue: 02 March - April 2026

ISSN 2582-0176

To determine the reliability of the structured checklist observation instruments (covering the implementation of teaching and learning strategies, use of teaching resources and materials, use of facilities and equipment, and implementation of evaluation), the researcher employed the Kuder-Richardson 21 (KR-21) calculation method. The Kuder-Richardson formula is as follows:

$$r^{21} = \frac{k\,(s^2) - X\,(k - X)}{(k - 1)\,(s^2)}$$

Petunjuk:

k- item
s^2 - sisihan piawai kuasa dua
X – min

The Kuder-Richardson method is similar to Cronbach's Alpha, differing only in the calculation of item variance to determine the reliability of an instrument. The Kuder-Richardson formula can be applied to instruments containing dichotomous items, where scores are assigned values such as 0 or 1. Specifically, Abdul Rashid & Siti Rahayah (1999) state that the KR-21 formula is used when it is assumed that every item has an equal level of difficulty or an identical correct/incorrect response structure. Both the Kuder-Richardson and Cronbach's Alpha formulas are used to measure the internal consistency of an instrument.

Inter-rater reliability, on the other hand, refers to the degree to which different observers produce the same results. Observations are conducted until an inter-rater reliability coefficient of 80% or higher is achieved. This reliability measure is used to evaluate the observation instrument related to the implementation of teaching and learning strategies. Inter-rater reliability is particularly relevant here because the data is obtained through observation. The formula for the inter-rater reliability coefficient is as follows:

$$\frac{Number\ of\ actual\ agreements}{Number\ of\ possible\ agreements} \text{ x } 100$$

The data collected through observations using the structured checklist in this study is guided by the established research objectives. A pilot study was conducted to determine the reliability of the instrument, as shown in Table 1. This pilot study was carried out in secondary schools within the Hulu Selangor district.

Table 1: Coefficients and Reliability Values for Observation (Structured Checklist).

Instrument	Variable / Technique	Analysis Technique	Reliability Value
Structured Checklist	Implementation of Teaching and Learning Strategies	K-R 21	0.782
	Interrater Reliability	Percentage Agreement	91.7%
	Use of Teaching Resources and Materials	K-R 21	0.985

Volume: 09, Issue: 02 March - April 2026

ISSN 2582-0176

Instrument	Variable / Technique	Analysis Technique	Reliability Value
	Use of Facilities and Equipment	K-R 21	0.600
	Implementation of Physical Education Assessment	K-R 21	0.819

NK – Pekali / Nilai Kebolehpercayaan

Meanwhile, the pilot study for the observation instrument in the form of a structured checklist was conducted among 30 Physical Education teachers in secondary schools in the Klang district. This pilot study was carried out with the assistance of Physical Education department teachers at the secondary school level. During the implementation of the pilot study, several pieces of feedback were received from the respondents, who consisted of teachers teaching Physical Education. Based on that feedback, the researcher updated the observation instrument (structured checklist).

After the instrument was updated, the researcher conducted the pilot study once more. The results from the second pilot study found that the reliability level of the observation instrument possessed a high reliability coefficient value. Based on the findings of the pilot study, the researcher also decided to use the observation instrument (structured checklist) as the method or technique for data collection in the actual study.

4.0 DISCUSSIONS

4.1 Instrument's Validity

The findings indicate that the structured checklist possesses a high level of content validity, resulting from the validation by experts in the field of physical education. The developed items successfully reflect the key aspects of the teaching process, such as the implementation of teaching strategies, the use of teaching resources and materials, the use of facilities and equipment, and the implementation of assessment. This is consistent with the view of Chacon-Moscoso et al. (2018), who emphasize the importance of content validity in sports observation studies. Face validity was also supported as teachers and observers rated the checklist as easy to use and relevant to the classroom and field contexts, aligning with the findings of McKenna & Thompson (2015), which stress the need for practical and clear observation instruments. This proves that the instrument truly measures the intended teaching components, in line with pedagogical theories and national curriculum standards.

4.2 Instrument's Reliability

From the perspective of reliability, the findings show a high inter-rater reliability value, indicating that different observers are capable of producing consistent scores when using the same checklist. A study by Van der Mars (1989) also emphasizes that inter-rater reliability is a critical aspect of physical education observation. This is vital because the observation process involves the interpretation of teacher and student behaviors, and consistency between observers ensures that the collected data is not influenced by individual bias. Additionally, intra-rater reliability demonstrated stability when the same observer re-evaluated at different times,

Volume: 09, Issue: 02 March - April 2026

ISSN 2582-0176

resulting in similar scores. These results confirm that the structured checklist possesses strong internal consistency, in line with the procedures outlined by DeMonbrun & Finelli (2015) in the development of observation protocols.

4.3 Implications for Teaching Implementation

The findings regarding validity and reliability carry significant implications for the implementation process of physical education teaching in secondary schools. First, the observation instrument can serve as a standardized tool to evaluate the quality of teacher instruction systematically. Second, it assists school authorities and administrators in identifying strengths and weaknesses in teaching implementation, thereby allowing for more precise interventions or professional training. Third, a valid and reliable structured checklist enhances the credibility of teacher action research and educational studies, as the findings are derived from a robust instrument.

4.4 Suggestions for Improvement

Although the instrument demonstrates good levels of validity and reliability, there are several items that may require refinement to increase sensitivity toward variations in teacher instructional styles. Future studies could evaluate construct validity more deeply through factor analysis to ensure that every teaching dimension is clearly represented. Furthermore, observation training should be strengthened to ensure more uniform use of the checklist, thereby maintaining a high level of reliability.

5.0 CONCLUSION

This study proves that the structured checklist observation instrument has satisfactory levels of validity and reliability in evaluating the implementation process of physical education teaching in secondary schools. High content and face validity indicate that the checklist items truly reflect relevant teaching components, while consistent inter-rater and intra-rater reliability affirm that this instrument can be used repeatedly without compromising the accuracy of the findings.

This instrument not only supports the systematic evaluation of teacher instructional quality but also serves as a tool for reflection and professional development. With a valid and reliable checklist, teachers can identify aspects of their teaching that need improvement, while school administrators can use the findings to plan more focused interventions. Overall, this study contributes to the field of physical education by providing an observation instrument that is robust, practical, and suitable for the secondary school context. This instrument has the potential to become a standard reference in teacher action research, program evaluation, and educational research that emphasizes the effectiveness of teaching and learning.

REFERENCES

1. Abdul Rashid & Siti Rahayah. (1999).Isu dalam penilaian pendidikan. Bangi: Universiti Kebangsaan Malaysia.

2. Adi, S., Nurharsono,T., Billiandri, B., & Soenyoto,T.(2020). Validity and reliability of instruments in physical education learning multimedia. Journal of Physical Education and Sports, 9(2), 123-132.
3. Aiken. (1997). Questionnaires and inventories. Surveying opinions and assessing personality. New York: John Wiley & Sons.
4. Alias (1999).Statistik penyelidikan dalam pendidikan dan sains sosial. Bangi: Universiti Kebangsaan Malaysia.
5. Armstrong, Gosling, Weinman, Marteau. (1997). The place of inter-rater reliability in qualitative research: An empirical study. Sociology, vol.31, No.3, 597-606
6. Azizi, Shahrin, Jamaludin, Yusof & Abdul Rahim. (2007).Menguasai penyelidikan dalam pendidikan. teori, analisis & interpretasi data. Kuala Lumpur: PTS.
7. Charles A. Peck., Maia Goodman Young, & Wenqi zhang (2021). Using Teaching Performance Assessments for Program Evaluation and Improvement in Teacher Education. Paper work. National Academy of Education Committee on Evaluating and Improving Teacher Preparation Programs.
8. D'Aoust R. (2025). Course evaluation instrumentation: Development and validation for alignment with teaching effectiveness and learner outcomes. Teaching and learning in Nursing. Volume 20, Issue 4, pages 326-332.
9. Balachandher .(2005). Data analysis. In Richardson, S.,Balachandher Krishnan Guru, Cheng Ming Yu., Khong Kok Wei., Leo Pointon (Eds). How to research. A guide for undergraduate & Graduate students. (p. 39-64). Malaysia: Thomson.
10. Baumgartner T.A., Jackson A.S., Mahar M.T., Rowe, D.A. (2003). Measurement for evaluation in Physical Education and Exercise Science. Edition 7, McGraw-Hill.New York/Boston.
11. Boyle & Fisher. (2007). Educational tetsing. A competence-Based Approach. Australi : BPS Blackwell.
12. Bhasah. (2007).Kaedah menjalanka kajian tindakan di sekolah. Kuala Lumpur : Miharja.
13. Chacon-Moscoso, S., Sanduvete-Chaves,S, Anguera,M.T., Losada,J.L., Portell,M.&Lozano-Lozano,J.A.(2013). Preliminary checklist for reporting observational studies in sports areas:Content validity. Frontiers in Psychology, 9, 291.
14. Cicciarella. (1997). Research in physical education, Exercise science, and sport. Arizona: Gorsuch Scarisbrick.
15. Cohen, Manion, & Morrison. (2007). Quantitative methods in psychology. A Power primer. Psychological Bulletin. Vol.112, No., 155 – 159. New York.
16. DeMonbrun,R.M.,& Finelli,C.J.(2015). Methods for establishing validity and reliability of observation protocols. Center for Research on Learning and Teaching in Engineering University of Michigan.
17. Gronlund. (2003). Assessment of student achievement. 7th ed. Boston: Pearson Education.
18. Gurvitch, BlankenshipMetzler, & Lund. (2008). Student teachers' implemented of Model-Based Instruction: Facilitators and Inhibitors. Journal of Teaching in Physical Education, 27, 466-486.
19. Harris,D.N.(2012). Classroom observation as a method for research and improvement. Carnegie Knowledge Network Brief. Retrieved from https://www.carnegieknowledgenetwork.org.

Volume: 09, Issue: 02 March - April 2026

ISSN 2582-0176

20. Krejcie & Morgan. (1970). Determining Sample Size for Research Activities. Educational and Psychologycal Measurement, 30, 607-610.
21. Medina M.S. et al. (2019). A Review of Strategies for Designing, Administratering, and Using Student Ratings of Instruction. Am J Pharm Educ. 83(5): 7177.
22. Mckenna, J.,&Thompson. (2015). Physical Education Teacher effectiveness. Validity and reliability of structured observation instruments. European Physical education Review, 21(3), 345 - 362.
23. Mohd Majid. (1994).Kaedah penyelidikan pendidikan. Kuala Lumpur: Dewan Bahasa dan Pustaka.
24. Mohnsen, B.S. (2003). Teaching middle school physical education. 2nd ed.USA: Human Kinetics.
25. Noraini, Loh, Ahmad Zabidi, Norjoharuddeen & Rahimi. (2006). Effective implementation of teaching mathematics and science in English: A case study. Jurnal Pendidikan, Fakulti Pendidikan, Universiti Malaya, 101 – 115.

