Original Article

Reliability and feasibility of an evaluation tool of teacher effectiveness and the quality of physical education sessions

AINA MARIA GALMES-PANADES[1,2,3] , PERE PALOU-SAMPOL[1,2,3] , PERE ANTONI BORRAS[1,2,3], ADRIA MUNTANER-MAS[1,2,3]

[1]*Faculty of Education, Department of Pedagogy and Specific Didactics, University of the Balearic Islands, Palma, Spain*
[2]*Research Group in Physical Activity and Sport Science, University of the Balearic Islands (GICAFE-UIB), Palma, Spain*
[3]*Institute for Educational Research and Innovation (IRIE), Palma, Spain*

ABSTRACT

The need to ensure quality physical education has motivated the development of a reliable method to evaluate physical education sessions, capable of detecting areas of improvement, thus facilitating the continuous growth of the teaching process. The main aim of the present study is to design a feasible and reliable evaluation instrument for physical education sessions, and to determine its reliability and feasibility with a group of practicing physical education teachers. Two evaluations of the same session (videotaped) were conducted to determine whether the instrument was reliable and feasible (n = 20), 9 women and 11 men, leaving 2 weeks between the first and the second evaluation. One-way analysis of variance (ANOVA) and intraclass correlation coefficient (ICC) were performed. The total score of the evaluation instrument shows an ICC of 0.953, (95%CI 0.881; 0.982), with a *p*-value ≤ .001. When comparing the responses of teachers from different educational levels (primary, secondary, and university), no significant differences were found between responses to any of the items or in the total score on either occasion. Total score in the test was (69.3 mean, [14.0 SD]), *p*-value = .881; and (66.2, [13.9]) *p*-value = .943 in the retest. In conclusion, the present evaluation instrument is feasible and reliable for the evaluation of physical education sessions.
Keywords: Numerical observational scale; Teacher evaluation; Primary education; Secondary education; University education.

Cite this article as:
Galmes-Panades, A.M., Palou-Sampol, P., Borras, P.A., & Muntaner-Mas, A. (2021). Reliability and feasibility of an evaluation tool of teacher effectiveness and the quality of physical education sessions. *Journal of Human Sport and Exercise, in press*. https://doi.org/10.14198/jhse.2023.181.14

Corresponding author. *Faculty of Education, Department of Pedagogy and Specific Didactics, University of the Balearic Islands, Carretera Valldemossa, km 7.5, Guillem Cifre Building. Palma 07120, Spain.* https://orcid.org/0000-0001-6977-9874
E-mail: aina.galmes.panades@gmail.com
Submitted for publication August 09, 2021.
Accepted for publication September 22, 2021.
Published *in press* November 17, 2021.
JOURNAL OF HUMAN SPORT & EXERCISE ISSN 1988-5202.

doi:10.14198/jhse.2023.181.14

INTRODUCTION

Teacher effectiveness in physical education and understanding that the educational system needs to provide quality physical education converged and resulted in the need to find a feasible and reliable method to assess teacher effectiveness in physical education. For this reason, researchers in the field called for the need to broaden science in physical education teacher effectiveness and physical educators' perceptions of effective teaching (Dyson, M., & Plunkett, 2014; Kyrgiridis, Derri, Emmanouilidou, Chlapoutaki, & Kioumourtzoglou, 2014; Lindsay, 2014; Mercier, K., & Doolittle, 2013; Metzler, 2014; J. Rink, 2014; J. E. Rink, 2013; Ward, 2013).

Quality physical education requires an adequate infrastructure to ensure the students' opportunity to learn, meaningful content, and appropriate pedagogical practices, including good classroom management and assessment of the students, the session and the teaching task (Fernández-Rivas & Espada-Mateos, 2019; National Association for Sport and Physical Education (NASPE), 2007; Navarro-Patón, Lago-Ballesteros, Basanta-Camiño, & Arufe-Giraldez, 2019). Likewise, quality physical education requires appropriate methods for assessing student learning and teacher effectiveness in physical education; Moreover, it also needs clearly defining objectives related to student outcomes in physical education (Dyson, M., & Plunkett, 2014; Kyrgiridis et al., 2014; Mercier, K., & Doolittle, 2013; Metzler, 2014; J. Rink, 2014; J. E. Rink, 2013; Ward, 2013). From a formative perspective, evaluation is a process that aims to improve the teaching-learning process, with a triple purpose: to improve student learning, to improve teaching practice, and to introduce improvements in the teaching-learning process during the same process, be it a didactic unit, a term, or an academic year (González Palacio, Chaverra Fernández, Bustamante Castaño, & Toro Suaza, 2020; López Pastor & Pérez Pueyo, 2017).

The evaluation instruments that have been used in recent years have become obsolete and there is a need to reformulate and simplify the evaluation instruments to adapt them to the characteristics of the current physical education sessions: high ratios, complex methodologies with a large number of interactions and the importance of feedback and formative assessment as a pillar of learning. Investigating teacher effectiveness in physical education using a reliable observation instrument benefits the discipline of physical education and, in addition, results in the development of active, and therefore healthier, lifelong citizens (Burgueño et al., 2019; Caracuel, Padial, Torres, & Cepero, 2020; Edufisaludable, 2021).

Effective teaching was related to features such as teacher preparation, lesson planning, content application, classroom organization and management, teaching strategies, positive learning environment, class control and discipline, teacher flexibility, communication skills, teacher feedback, and evaluation (Fernández, Hortigüela, & Pérez, 2018). To assess these characteristics in a physical education session, a wide range of evaluation instruments can be used, among which the following can be highlighted: checklists, verbal observation scales, numerical scales, descriptive scales or rubrics, graphic scales, individual monitoring sheets, group monitoring sheets, questionnaires, student's notebook, teacher's notebook, etc. (Hamodi, López Pastor, & López Pastor, 2015).

Currently, physical education seeks to promote the active participation of students, encouraging their autonomy, decision-making, creativity, and a high level of motor engagement during the sessions. It also seeks to give the teacher a reflective role in the teaching-learning process and encourages self-evaluation and peer evaluation, as well as heteroevaluation by the students, which lead to the continuous improvement of the teaching-learning process (González Palacio et al., 2020).

Self-evaluation is the evaluation carried out by the person themselves and refers to both the student and the teacher (García Carrillo, 2016). Self-evaluation is a fundamental element in the training process, through which teachers work on autonomy, criticism, and reflection to review, assess and pass judgment on their actions. As teachers, continuous training, reflection, and self-criticism, among others, are fundamental aspects to guarantee the quality of the teaching-learning process. Peer evaluation or co-evaluation is the joint evaluation where the development of peers is assessed, and offers the opportunity to get to know one's own potential, as well as giving room for improvement (Borjas, 2011).

This study focuses on improving the teaching practice by developing an evaluation instrument for physical education sessions, from which teachers can identify strengths and areas for improvement. The main aim of the present study is to design a feasible and reliable evaluation instrument for physical education sessions.

MATERIAL AND METHODS

Within the framework of the compulsory subject of Physical Education and Healthy Habits, in the primary education degree of the Balearic Islands University, during the 2020-2021 school year, a teaching innovation project was carried out to design a feasible and reliable evaluation instrument for physical education sessions.

Participants

For the design, and to assure the feasibility and reliability of this evaluation instrument, 56 students and 23 teachers and professors participated. In the first phase of the design of the evaluation instrument, the sample consisted of 56 third-year students of the University of the Balearic Islands. The students were between 21 and 40 years old, 40 (71.4%) were women and 16 (28.6%) men. All the students lived in Mallorca, Spain, and attended the face-to-face sessions of the course every fortnight. Inclusion criteria were that the students had to be enrolled in the Physical Education and Healthy Habits subject, and to be taking the classroom-based itinerary.

In the second phase of the design, the sample consisted of a group of 3 external experts. One expert was a primary school physical education teacher, the other was a secondary school physical education teacher and the third was a university physical education teacher, two men, and one woman, respectively.

In the third phase, the sample consisted of a group of 20 active physical education teachers, 8 from primary education, 6 from secondary education, and 6 from university, of which 9 (45%) were women and 11 (55%) men. All teachers lived in Mallorca, Spain. Teachers were selected from a list of teachers collaborating with the university department of pedagogy and specific didactics. The inclusion criteria where they had to be active workers at the time of the assessments, and they had to wish to participate in this study voluntarily (See Figure 1. Flow chart).

Measures

To carry out this innovation project, the university students were previously informed of the characteristics of the project and its relationship with the subject. The practical sessions of the subject were designed to be teaching simulations. In working groups, the students played the role of teachers, while the rest of the students simulated being primary school pupils.

The first phase - design: In each practical session of the subject, a working group evaluated the session of the group that was the speaker, following the numerical observation scale designed by the teachers, based on a literature review. Once the observation scale had been completed, the evaluation form was sent to the

teachers of the subject. The teacher effectiveness evaluation instrument included two questions for students to answer: "*Did you miss any item to evaluate the session? If so, please indicate which one*" and "*Did you find any item not relevant to evaluate the session? If so, please indicate which one*". Based on the students' responses, the instrument was revised again to adapt it better to the needs and characteristics of the evaluation of physical education. The 56 students used the evaluation instrument a total of 76 times during the course.

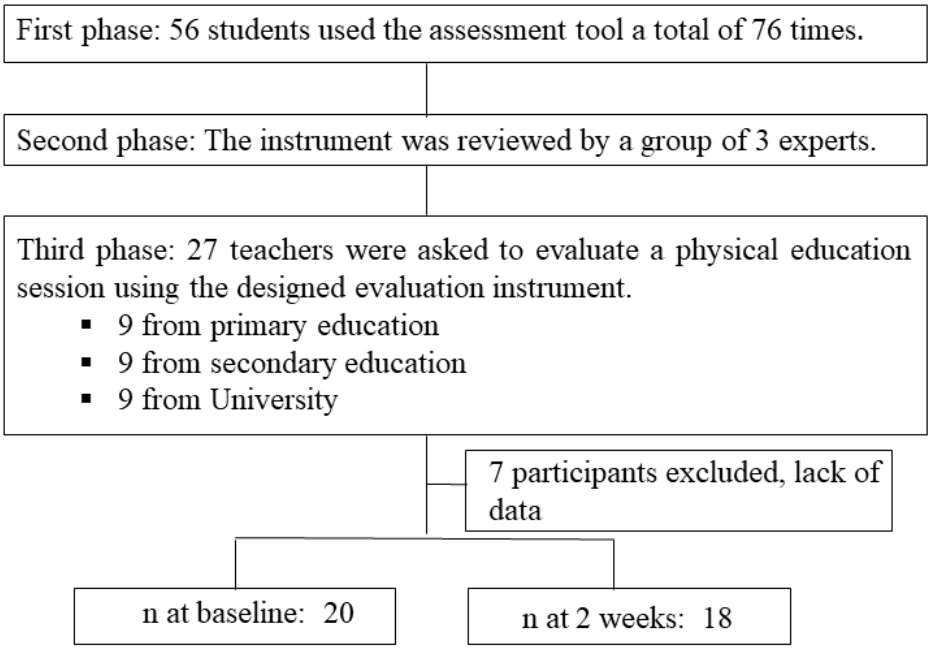

Figure 1. Flow of participants in the design and reliability of the assessment tool for physical education sessions.

The second phase - design: After reviewing the students' comments and making the relevant adaptations, as well as a second literature review, the evaluation instrument was sent to a group of 3 external experts, who evaluated the concreteness, appropriateness, and comprehension of each item on a scale of 1 to 3, where 1 was inadequate; 2 was adequate, and 3 was highly satisfactory. Based on the feedback from the group of experts, changes were incorporated into the evaluation instrument.

The third phase – feasibility and reliability: Once the final version of the evaluation instrument was developed, a group of primary, secondary, and university teachers were sent a recording of a primary school physical education session and the evaluation instrument so that they could view the recording of the session and evaluate it using the instrument. After 2 weeks, the group of teachers viewed the same session again and evaluated it using the evaluation instrument.

Procedures

Before the start of the course, a literature review was carried out on which the design of the evaluation instrument for the physical education sessions was based (Cano García, Pons Seguí, & Lluch Molins, 2018; Danielson, 2014; Villarroel & Bruna, 2019), including the Society of Health and Physical Educators of America (SHAPE) recommendations for effective physical education (SHAPE America — Society of Health and Physical Educators, 2016) and the recommendations of the National Association for Sport and Physical

Education (NASPE) (National Association for Sport and Physical Education (NASPE), 2007), which sets standards for quality physical education programs. In addition to the main recommendations of these bodies, other aspects considered relevant were taken into account, such as the inclusion of a specific item to assess co-education during the session (Sánchez Torrejón & Barea Villalba, 2019). The evaluation instrument was a numerical observation scale, consisting of 24 items and 5 categories: 1: Unsatisfactory, unacceptable, does not meet standards, needs revision; 2: Inadequate, below average, needs improvement, underdeveloped; 3: Good, satisfactory, basic, meets requirements; 4: Very good, above average, competent; 5: Exemplary, clearly exceeds standards, mastery (first phase).

The instrument also includes a section for free observations, where teachers can make any annotation they consider appropriate. The instrument was designed considering the evaluation of 3 main axes: the management of spaces, materials, and groupings; the tasks; and the teaching intervention. In addition to the relevance, efficiency, and effectiveness of the instrument to evaluate the physical education sessions, another aspect that was given special importance during the design was the convenience of the instrument to be applied during the sessions, so it had to be brief, collect the information visually, and be easy to fill in (first phase).

Analysis

The first and second phases of the study were used to design the evaluation instrument, while the third phase was used to implement the evaluation instrument and analyse its reliability and feasibility.

Characteristics of the study participants during the third phase were presented as absolute numbers (percentages) for categorical variables.

Differences between the first and second evaluations from the teachers' group were presented as mean and standard deviations (SD). One-way analysis of variance (ANOVA) was used to assess differences between the first and second evaluation across categories of education level (primary, secondary, and university). For the test reliability and feasibility, the intraclass correlation coefficient (ICC) was calculated between test and retest results (third phase). The interpretation of ICC values was based on standardized guidelines, in which a value less than 0.5 indicates poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability (Altman & Bland, 1983; Muntaner-Mas, Martinez-Nicolas, Quesada, Cadenas-Sanchez, & Ortega, 2021). All analyses were conducted with the Stata v13.0. program and SPSS v24 program. All p-values < .05 were deemed statistically significant.

RESULTS

The main result of this research is the reliability and feasibility of the evaluation instrument for evaluating physical education sessions, being reliable for teachers to use it at all 3 levels of education: primary education, secondary education, and university. As a result of the third phase of the present study, in which a group of teachers from the 3 educational levels used the evaluation instrument on 2 occasions, it was detected that in some cases, teachers found it difficult to evaluate some aspects of the session, because they could not observe them accurately. For this reason, it was decided to incorporate a sixth response option to evaluate each item. In addition to a rating on a scale of 1 to 5, where 1 was "*unsatisfactory*" and 5 was "*exemplary*", a "*not applicable*" option was included for those situations where, due to the nature of the session, the item could not be observed. An example of this case would be item 3 "*Manages material during the session to maximize practice time*" if, during the session being assessed, the teacher does not use any

material. The instrument is presented below in English (see Table 1), in addition to including the instrument translated into Spanish and Catalan (see Annexes 1 and 2).

Table 1. Numerical observation scale for the evaluation of physical education sessions.

EVALUATION OF THE TEACHING PROCESS IN PHYSICAL EDUCATION	
*1. Unsatisfactory, unacceptable, does not meet standards, needs revision. 2. Inadequate, below average, needs improvement, underdeveloped. 3. Good, satisfactory, basic, meets requirements. 4. Very good, above average, competent. 5. Exemplary, clearly exceeds standards, mastery. NA. Not applicable: For situations where, due to the nature of the session, the item cannot be observed.	**Num.**
SPACES, MATERIALS AND GROUPINGS	
1. Limits unnecessary **shifts** between tasks by favouring quick transitions.	
2. Prepares the **material** and its distribution **in advance** of the session and with efficiency criteria.	
3. Manages **material during** the session to maximise practice time.	
4. **Groups and organises** the students so that they are always active, and no one is standing still.	
TASKS	
5. Adequately schedules the **number of tasks and the time** devoted to each one in order to achieve the learning objectives, contents and key competences.	
6. Designs tasks to develop **motor competence**.	
7. Designs tasks to encourage maximum **participation** of all learners and motor engagement time (equal to or greater than 50% of the session).	
8. Ensures the practical **participation** of all learners, irrespective of their level of motor competence.	
9. Assigns **tasks** to **students** who **cannot perform** the **practical session** (injuries, inappropriate clothing, etc.).	
10. Ensures the **safety** of students in all tasks.	
INTERVENTION	
11. Explains the **objectives** of the session at the beginning and relates them to previous learning. They are worked on through the 3 areas: cognitive, psychomotor and affective.	
12. **Communicates** clearly, provides sufficient and well-structured information.	
13. **Exemplifies** and resolves doubts before starting the task.	
14. **Captures students' attention** and develops listening routines.	
15. Shows **mastery of the content** taught.	
16. Gives **feedback** that motivates learners.	
17. **Moves around** the space to observe, listen, give feedback, resolve doubts, etc.	
18. **Adapts** to unforeseen events, changing situations and students' needs.	
19. **Listens** to the students and creates a climate of trust that favours communication.	
20. **Manages time** efficiently throughout the session, devoting sufficient time to each activity and following the timetable.	
21. Takes into account **co-education** throughout the session: language, contents and methodologies.	
22. Records aspects of the session for **learner evaluation**.	
23. Records aspects of the session for the **evaluation** of the **session**.	
24. Concludes the session with a **group reflection**, giving feedback and motivating the learners.	
OBSERVATIONS:	

Twenty-seven participants were included in the third phase for field-test reliability and feasibility. A total of 7 participants were excluded from the test and retest, due to the lack of data, and 2 participants were lost to follow-up in the retest. Therefore, a total of 20 participants, 11 men, and 9 women were included in the analysis (phase 3). Descriptive characteristics of the participants are shown in Table 2.

Table 2. Characteristics of the teachers who carried out the evaluation in the third phase (n = 20).

Characteristics		**Number (%)**
Educational level	Primary education	8 (40)
	Secondary education	6 (30)
	University	6 (30)
Sex	Men	11 (55)
	Women	9 (45)

Note: Values are number (percentage) for categorical variables.

Table 3 shows a comparison of test-retest responses, item by item, to determine the reliability of the instrument (third phase). Most of the ICCs were greater than 0.7, indicating moderate-good reliability, and 6 of the 24 items had an ICC greater than 0.90 indicating excellent reliability. All items except 1 showed a significant correlation in the test-retest. Furthermore, the total score of the evaluation instrument had an ICC of 0.953, (95% CI 0.881; 0.982), $p \leq .001$, indicating excellent reliability of the instrument.

Table 3. Item-by-item reliability measured by comparing test and retest responses.

	Total n	**ICC**	**(95% CI)**	***p*-value**
Item 1	18	0.864	(0.678; 0.947)	<.001
Item 2	18	0.869	(0.689; 0.948)	<.001
Item 3	18	0.939	(0.843; 0.977)	<.001
Item 4	18	0.851	(0.650; 0.941)	<.001
Item 5	18	0.792	(0.526; 0.917)	<.001
Item 6	18	0.922	(0.806; 0.970)	<.001
Item 7	18	0.874	(0.700; 0.951)	<.001
Item 8	18	0.752	(0.436; 0.900)	<.001
Item 9	18	0.029	(-0.471; 0.491)	.456
Item 10	18	0.927	(0.816; 0.972)	<.001
Item 11	18	0.864	(0.675; 0.947)	<.001
Item 12	18	0.645	(0.262; 0.851)	.002
Item 13	18	0.702	(0.371; 0.876)	<.001
Item 14	18	0.903	(0.759; 0.963)	<.001
Item 15	18	0.901	(0.757; 0.962)	<.001
Item 16	18	0.933	(0.829; 0.974)	<.001
Item 17	18	0.807	(0.556; 0.923)	<.001
Item 18	18	0.818	(0.551; 0.933)	<.001
Item 19	18	0.788	(0.523; 0.914)	<.001
Item 20	18	0.831	(0.583; 0.935)	<.001
Item 21	18	0.557	(0.122; 0.810)	.008
Item 22	18	0.779	(0.497; 0.912)	<.001
Item 23	18	0.779	(0.497; 0.912)	<.001
Item 24	18	0.458	(-0.028; 0.765)	.032
Total	**18**	**0.953**	**(0.881; 0.982)**	**<.001**

The values show the ICC (95% CIs), model two-way mixed and single measures. These represent the correlation between test and retest responses, analysed item-by-item and the total score. Analyses included only completers. Abbreviations: ICC: intraclass correlation coefficient; CI: confidence interval.

Table 4 shows a comparison of teachers, according to the educational level at which they teach (primary, secondary, and university), concerning to the answers given to each item in the test and retest (third phase). No significant differences were found between the teachers' responses to any of the items or in the total score on either occasion. Total score in the test was (69.3 mean, [14.0 SD]), *p*-value = .881; and (66.2, [13.9]) *p*-value =.943 in the retest.

Table 4. Test and re-test results across categories of education level.

	Test			**Retest**		
	Total n	Mean (SD)	*p*-value	Total n	Mean (SD)	*p*-value
Item 1	20	3.10 (0.97)	.584	18	3.17 (1.04)	.617
Item 2	20	3.50 (0.83)	.212	18	3.50 (1.04)	.477
Item 3	20	3.60 (0.94)	.675	18	3.56 (0.92)	.797
Item 4	20	2.90 (1.07)	.965	18	2.72 (1.02)	.731
Item 5	20	2.65 (0.93)	.838	18	2.56 (0.78)	.516
Item 6	20	2.90 (0.85)	.945	18	2.72 (0.83)	.866
Item 7	20	3.45 (1.15)	.610	18	3.11 (1.02)	.446
Item 8	20	3.35 (0.88)	.439	18	3.00 (0.84)	.805
Item 9	20	1.30 (0.57)	.596	18	1.22 (0.55)	.836
Item 10	20	3.35 (0.81)	.791	18	3.33 (0.91)	.648
Item 11	20	2.00 (0.73)	.296	18	2.00 (0.84)	.376
Item 12	20	3.20 (0.77)	.504	18	3.00 (0.84)	.822
Item 13	20	3.10 (0.97)	.866	18	2.83 (0.86)	.522
Item 14	20	3.20 (0.76)	.886	18	3.11 (0.76)	.293
Item 15	20	3.20 (0.89)	.915	18	3.06 (0.94)	.930
Item 16	20	2.90 (0.91)	.396	18	2.78 (0.94)	.207
Item 17	20	3.35 (0.75)	.315	18	3.22 (0.94)	.487
Item 18	18	2.83 (0.92)	.995	18	2.78 (0.81)	.465
Item 19	20	3.75 (0.91)	.956	18	3.50 (0.92)	.192
Item 20	20	3.05 (1.19)	.977	18	2.61 (0.98)	.559
Item 21	20	3.15 (0.75)	.476	18	3.06 (0.87)	.274
Item 22	20	1.45 (0.60)	.454	18	1.33 (0.59)	.251
Item 23	20	1.45 (0.60)	.454	18	1.33 (0.59)	.251
Item 24	19	3.00 (1.11)	.647	18	2.67 (1.24)	.407
Total	**20**	**69.3 (14.0)**	**.881**	**18**	**66.2 (13.9)**	**.943**

Values are mean (SD) for continuous variables. The mean (SD) corresponds to the total number of teachers who took the evaluation, while the p-value shown is the comparison between the responses of teachers at the 3 levels of education: primary, secondary and university. Each item could receive a score from 1 to 5. The total score could vary between 24 and 120. The description corresponding to each numerical value was as follows: 1. Unsatisfactory, unacceptable, does not meet standards, needs revision; 2. Inadequate, below average, needs improvement, underdeveloped; 3. Good, satisfactory, basic, meets requirements; 4. Very good, above average, competent; 5. Exemplary, clearly exceeds standards, mastery.

Furthermore, a t-test was performed to determine differences by sex (third phase), and no statistically significant differences were found on none of the occasions, neither in the test nor in the retest (*p*-value = .774 and .518, respectively), (data not shown).

DISCUSSION

Results from this study show that the evaluation instrument for physical education sessions is reliable and feasible. The evaluation instrument will be used to evaluate the physical education sessions and the teaching

task. It is therefore a very useful instrument to encourage reflection and improvement of the teaching-learning process.

Like the instrument of the NAPSE (National Association for Sport and Physical Education (NASPE), 2007), the aim of this evaluation instrument is, in addition to assisting principals, school district curriculum specialists, and others who evaluate physical education teachers; to provide guidance to physical education teachers for co-evaluation and reflection, as well as to serve as an instructional instrument in university physical education teacher education programs. In addition, the evaluation instrument is presented in 3 languages: English, Spanish, and Catalan, to reach as many teachers as possible with this instrument.

Even though the session visualised in the third phase, to be evaluated by the evaluation instrument designed, was a primary school session, no differences were found between the teachers of the 3 educational levels, nor between men and women. We are aware that more studies are needed to validate the instrument to evaluate physical education sessions at the three levels, but we consider that the results obtained are very promising and that the instrument is very simple and practical to use in day-to-day physical education sessions, which is an advantage that adds to the reliability and feasibility demonstrated through this study.

Furthermore, this instrument takes into account international recommendations for the evaluation of physical education (National Association for Sport and Physical Education (NASPE), 2007; SHAPE America — Society of Health and Physical Educators, 2016), but at the same time incorporates new aspects that allow it to be better adjusted to the current context, such as the inclusion of a sixth response option of "*not observed*" to adapt to the characteristics of each session, or the item that evaluates gender equality, evaluating and promoting a fundamental aspect of 21st-century education, co-education (González-Valero, San Román-Mata, Ubago-Jiménez, & Puertas-Molero, 2020; Sánchez Torrejón & Barea Villalba, 2019).

A marked strength of this study was the methodology used in each phase, which has made it possible to assess the usefulness of the instrument on 3 occasions: with the students of the primary education degree (first phase), with the group of experts (second phase), and with the group of primary, secondary and university teachers (third phase). It is important to highlight the innovation of the project, incorporating very relevant aspects into the instrument to achieve the education we want for the next generations: a participatory, responsible, healthy, inclusive, and equal population. Finally, the consistency of the results shows the reliability and feasibility of the designed instrument. In terms of limitations, the present study has not validated the instrument. Although the sample size was limited, conclusive results have been obtained. Using a single physical education session to measure feasibility and reliability has limited the extrapolation of the results to educational levels other than primary education, although no significant differences were found in the responses of teachers at the 3 educational levels.

CONCLUSIONS

In conclusion, the presented instrument is feasible and reliable for evaluating physical education sessions in primary education. Moreover, it is relevant, visual, easy, and comfortable to use, which are very important characteristics for an evaluation instrument in physical education. Although no secondary education or university sessions were evaluated, no significant differences were found between the teachers of different levels who used the evaluation instrument. Further studies are needed to validate the present instrument.

AUTHOR CONTRIBUTIONS

AMGP, PPS, PAB and AMM conceived of the study. AMGP and AMM completed the statistical analysis. AMGP, PPS, PAB and AMM drafted the manuscript. AMGP supervised the study. All authors were involved in oversight of recruitment, data collection, revision of the manuscript and read and approved the final manuscript.

SUPPORTING AGENCIES

Activity carried out with the support of the IRIE. Call for grants for innovation and improvement of teaching quality. PID202103. Institute for Research and Innovation in Education (IRIE), University of the Balearic Islands, Spain.

DISCLOSURE STATEMENT

The authors declare that they have no competing interests. Authors are agreeing to Copyright Notice as part of the submission process.

ETHICS DECLARATIONS

All experimentation was conducted in conformity with ethical and humane principles of research.

ACKNOWLEDGEMENTS

The authors would like to thank the students and teaching staff for their participation and the Institute for Research and Innovation in Education (IRIE) of the University of the Balearic Islands for their funding, as well as to Ms. Neus Palou Ferrer for the linguistic revision.

REFERENCES

Altman, D. G., & Bland, J. M. (1983). Measurement in Medicine: The Analysis of Method Comparison Studies. The Statistician, 32(3), 307. https://doi.org/10.2307/2987937

Borjas, M. (2011). Peer assessment as a democratizing experience: Case of a Teacher Preparation Program. Revista Del Instituto de Estudios En Educación Universidad Del Norte.

Burgueño, R., González-Cutre, D., Sevil-Serrano, J., Herrador-Colmenero, M., Segura-Díaz, J. M., Medina-Casaubón, J., & Chillón, P. (2019). Understanding the motivational processes involved in adolescents' active commuting behaviour: Development and validation of the Behavioural Regulation in Active Commuting to and from School (BR-ACS) Questionnaire. Transportation Research Part F: Traffic Psychology and Behaviour, 62, 615-625. https://doi.org/10.1016/j.trf.2019.02.016

Cano García, M. E., Pons Seguí, L., & Lluch Molins, L. (2018). Análisis de experiencias de innovación docente universitarias sobre evaluación. Profesorado, Revista de Currículum y Formación Del Profesorado, 22(4). https://doi.org/10.30827/profesorado.v22i4.8392

Caracuel, R., Padial, R., Torres, B., & Cepero, M. (2020). The influence of healthy habits acquired by school-age adolescents in relation to physical education classes. In Journal of Human Sport and Exercise - 2020 - Summer Conferences of Sports Science. Universidad de Alicante. https://doi.org/10.14198/jhse.2020.15.Proc4.02

Danielson, C. (2014). The framework for teaching: Evaluation instrument. Journal of Research in Music Education (Vol. 66). https://doi.org/10.1177/0022429418793645

Dyson, M., & Plunkett, M. (2014). Enhancing interpersonal relationships in teacher education through the development and practice of reflective mentoring. (Interperso). https://doi.org/10.1007/978-94-6209-701-8_4

Edufisaludable. (2021). Why is it necessary to research physical education and the promotion of healthy habits? Retrieved from https://edufisaludable.com/en/why-is-it-necessary-to-research-physical-education-and-the-promotion-of-healthy-habits/

Fernández-Rivas, M., & Espada-Mateos, M. (2019). The knowledge, continuing education and use of teaching styles in Physical Education teachers. Journal of Human Sport and Exercise, 14(1). https://doi.org/10.14198/jhse.2019.141.08

Fernández, J., Hortigüela, D., & Pérez, Á. (2018). Revisando los modelos pedagógicos en educación física . Ideas clave para incorporarlos al aula. Revista Española de Educación Física y Deportes, (423), 57-80.

García Carrillo, L. S. (2016). La autoevaluación de los estudiantes un proceso por resignificar y reconstruir en la educación física escolar. Paideia Surcolombiana, (21), 27-42. https://doi.org/10.25054/01240307.1454

González-Valero, G., San Román-Mata, S., Ubago-Jiménez, J. L., & Puertas-Molero, P. (2020). Physical-healthy and psychosocial differences in school children: A study of gender. Journal of Human Sport and Exercise, 17(2). https://doi.org/10.14198/jhse.2022.172.07

González Palacio, E. V., Chaverra Fernández, B. E., Bustamante Castaño, S. A., & Toro Suaza, C. A. (2020). Diseño y validación de un cuestionario sobre las concepciones y percepción de los estudiantes sobre la evaluación en Educación Física (Design and validation of a questionnaire about the conceptions and perception of the students about the assessment in P. Retos, 2041(40), 317-325. https://doi.org/10.47197/retos.v1i40.80914

Hamodi, C., López Pastor, V. M., & López Pastor, A. T. (2015). Medios, técnicas e instrumentos de evaluación formativa y compartida del aprendizaje en educación superior. Perfiles Educativos, 37 (147), 146-161. https://doi.org/10.22201/iisue.24486167e.2015.147.47271

Kyrgiridis, P., Derri, V., Emmanouilidou, K., Chlapoutaki, E., & Kioumourtzoglou, E. (2014). Development of a Questionnaire for Self-Evaluation of Teacher Effectiveness in Physical Education (SETEQ-PE). Measurement in Physical Education and Exercise Science, 18(2), 73-90. https://doi.org/10.1080/1091367X.2013.866557

Lindsay, E. (2014). Effective teaching in physical education: The view from a variety of trenches. Research Quarterly for Exercise and Sport, 85, 31 -37. https://doi.org/10.1080/02701367.2014.873330

López Pastor, V. M., & Pérez Pueyo, Á. (2017). Evaluación formativa y compartida en educación experiencias de éxito en todas las etapas educativas. (U. de León, Ed.).

Mercier, K., & Doolittle, S. (2013). Assessing student achievement in physical education for teacher evaluation. Journal of Physical Education, Recreation & Dance, 84(3), 38-42. https://doi.org/10.1080/07303084.2013.767721

Metzler, M. (2014). Teacher effectiveness research in physical education: The future isn't what it used to be. Research Quarterly for Exercise and Sport, 85, 14 -19. https://doi.org/10.1080/02701367.2014.872932

Muntaner-Mas, A., Martinez-Nicolas, A., Quesada, A., Cadenas-Sanchez, C., & Ortega, F. B. (2021). Smartphone App (2kmFIT-App) for Measuring Cardiorespiratory Fitness: Validity and Reliability Study. JMIR MHealth and UHealth, 9(1), e14864. https://doi.org/10.2196/14864

National Association for Sport and Physical Education (NASPE). (2007). Physical education teacher evaluation tool. Retrieved April 16, 2021, from https://www.michigan.gov/documents/mde/NASPETool_212381_7.pdf

Navarro-Patón, R., Lago-Ballesteros, J., Basanta-Camiño, S., & Arufe-Giraldez, V. (2019). Relation between motivation and enjoyment in physical education classes in children from 10 to 12 years old. Journal of Human Sport and Exercise, 14(3). https://doi.org/10.14198/jhse.2019.143.04

Rink, J. (2014). Teacher Effectiveness in Physical Education-Consensus? Research Quarterly for Exercise and Sport, 85(3), 282-286. https://doi.org/10.1080/02701367.2014.932656

Rink, J. E. (2013). Measuring teacher effectiveness in physical education. Research Quarterly for Exercise and Sport, 84, 407-418. https://doi.org/10.1080/02701367.2013.844018

Sánchez Torrejón, B., & Barea Villalba, Z. (2019). Towards a violet school: Teachers' of Primary education initial training in coeducation. Tendencias Pedagógicas. https://doi.org/10.15366/tp2019.34.007

SHAPE America - Society of Health and Physical Educators. (2016). 20 Indicators of Effective Physical Education Instruction. Retrieved April 16, 2021, from https://www.shapeamerica.org/events/upload/20-Indicators-Brochure-WEB-003-2.pdf

Villarroel, V., & Bruna, D. (2019). ¿Evaluamos lo que realmente importa? el desafío de la evaluación auténtica en educación superior. Calidad En La Educación, (50), 492. https://doi.org/10.31619/caledu.n50.729

Ward, P. (2013). The role of content knowledge in conceptions of teaching effectiveness in physical education. Research Quarterly in Physical Education, 84, 431-440. https://doi.org/10.1080/02701367.2013.844045

ANNEX 1. Numerical observation scale for the evaluation of physical education sessions in Spanish.

EVALUACIÓN DEL PROCESO DE ENSEÑANZA EN EDUCACIÓN FÍSICA	
*1. Insatisfactorio, inaceptable, no cumple los estándares, necesita revisión. 2. Insuficiente, por debajo de la media, necesita mejora, poco desarrollado. 3. Bien, satisfactorio, básico, cumple los requisitos. 4. Muy bien, superior a la media, competente. 5. Ejemplar, supera claramente los estándares, maestría NP. No procede: Para aquellas situaciones en las que, por la naturaleza de la sesión, el ítem no se pueda observar	**Núm.**
ESPACIOS, MATERIALES Y AGRUPACIONES	
1. Limita los **desplazamientos** innecesarios entre tareas favoreciendo transiciones rápidas.	
2. Prepara el **material** y su distribución con **antelación** a la sesión y con criterio de eficacia.	
3. Gestiona el **material durante** la sesión para maximizar el tiempo de práctica.	
4. **Agrupa y organiza** al alumnado buscando siempre que estén activos y ninguno esté parado.	
TAREAS	
5. Programa adecuadamente el **número de tareas y el tiempo** dedicado a cada una para lograr los aprendizajes: objetivos, contenidos y competencias clave.	
6. Diseña las tareas para desarrolla la **competencia motriz**.	
7. Diseña las tareas para fomentar la máxima **participación** de todo el alumnado y el tiempo de compromiso motor (es igual o superior al 50% de la sesión).	
8. Asegura la **participación** práctica de todo el alumnado, con independencia de su nivel de competencia motriz.	
9. Asigna tareas al **alumnado** que **no** puede **realizar** la **sesión** práctica (lesiones, ropa inapropiada, etc.).	
10. Garantiza la **seguridad** del alumnado en todas las tareas.	
INTERVENCIÓN	
11. Explica los **objetivos** de la sesión a su inicio y los relaciona con aprendizajes anteriores. Se trabajan a través de los 3 ámbitos: cognitivo, psicomotor y afectivo.	
12. **Comunica** con claridad, aporta información suficiente y bien estructurada.	
13. **Ejemplifica** y resuelve las dudas antes de empezar la tarea.	
14. **Capta la atención** del alumnado y desarrolla rutinas de escucha.	
15. Muestra **dominio del contenido** que enseña.	
16. Ofrece **feedback** que motiva al alumnado.	
17. Se **desplaza** por el espacio para: observar, escuchar, dar feedback, resolver dudas, etc.	
18. Se **adapta** a imprevistos, situaciones cambiantes y necesidades del alumnado.	
19. **Escucha** al alumnado y crea un clima de confianza que favorece la comunicación.	
20. **Gestiona el tiempo** de forma eficiente a lo largo de la sesión, dedicando tiempo suficiente a cada actividad y siguiendo la programación.	
21. Tiene en cuenta la **coeducación** durante toda la sesión: lenguaje, contenidos y metodologías.	
22. Registra aspectos de la sesión para la **evaluación** del **alumnado**.	
23. Registra aspectos de la sesión para la **evaluación** de la **sesión**.	
24. Finaliza la sesión con una **reflexión grupal**, dando feedback y motivando al alumnado.	
OBSERVACIONES:	

ANNEX 2. Numerical observation scale for the evaluation of physical education sessions in Catalan.

EVALUACIÓ DEL PROCES D'ENSEÑANZA A EDUCACIÓ FÍSICA	
*1. Insatisfactori, inacceptable, no compleix els estàndards, necessita revisió. 2. Insuficient, per sota de la mitjana, necessita millora, poc desenvolupat. 3. Bé, satisfactori, bàsic, compleix els requisits. 4. Molt bé, superior a la mitjana, competent. 5. Exemplar, supera clarament els estàndards, mestratge NP. No procedeix: Per a aquelles situacions en les quals, per la naturalesa de la sessió, l'ítem no es pugui observar	**Núm.**
ESPAIS, MATERIALS I AGRUPACIONS	
1. Limita els **desplaçaments** innecessaris entre tasques afavorint transicions ràpides.	
2. Prepara el **material** i la seva distribució amb **antelació** a la sessió i amb criteri d'eficàcia.	
3. Gestiona el **material durant** la sessió per a maximitzar el temps de pràctica.	
4. **Agrupa i organitza** a l'alumnat buscant sempre que estiguin actius i cap estigui parat.	
TASQUES	
5. Programa adequadament el **nombre de tasques i el temps** dedicat a cadascuna per a aconseguir els aprenentatges: objectius, continguts i competències clau.	
6. Dissenya les tasques per a desenvolupa la **competència motriu.**	
7. Dissenya les tasques per a fomentar la màxima **participació** de tot l'alumnat i el temps de compromís motor (és igual o superior al 50% de la sessió).	
8. Assegura la **participació** pràctica de tot l'alumnat, amb independència del seu nivell de competència motriu.	
9. Assigna tasques a l'**alumnat** que **no pot realitzar** la sessió **pràctica** (lesions, roba inapropiada, etc.).	
10. Garanteix la **seguretat** de l'alumnat en totes les tasques.	
INTERVENCIÓ	
11. Explica els **objectius** de la sessió al seu inici i els relaciona amb aprenentatges anteriors. Es treballen a través dels 3 àmbits: cognitiu, psicomotor i afectiu.	
12. **Comunica** amb claredat, aporta informació suficient i ben estructurada.	
13. **Exemplifica** i resol els dubtes abans de començar la tasca.	
14. **Capta l'atenció** de l'alumnat i desenvolupa rutines d'escolta.	
15. Mostra **domini del contingut** que ensenya.	
16. Ofereix **feedback** que motiva a l'alumnat.	
17. Es **desplaça** per l'espai per a: observar, escoltar, donar feedback, resoldre dubtes, etc.	
18. **S'adapta** a imprevistos, situacions canviants i necessitats de l'alumnat.	
19. **Escolta** a l'alumnat i crea un clima de confiança que afavoreix la comunicació.	
20. **Gestiona el temps** de manera eficient al llarg de la sessió, dedicant temps suficient a cada activitat i seguint la programació.	
21. Té en compte la co-educació durant tota la sessió: llenguatge, continguts i metodologies.	
22. Registra aspectes de la sessió per a l'**avaluació** de l'**alumnat.**	
23. Registra aspectes de la sessió per a l'**avaluació** de la **sessió**.	
24. Finalitza la sessió amb una **reflexió grupal**, donant feedback i motivant a l'alumnat.	
OBSERVACIONS:	