THE IMPORTANCE OF DIFFERENT EVALUATION METHODS IN PHYSICAL EDUCATION – A CASE STUDY OF STRADDLE VAULT OVER THE BUCK

Matej Majerič, Janko Strel, Marjeta Kovač

Faculty of Sport, University of Ljubljana, Ljubljana, Slovenia

Original article

Abstract

This article presents a study of three different evaluation methods for one of the most frequently evaluated skills in physical education: straddle vault over the buck. The sample of measured subjects included 193 13-year-old boys, whose video-recorded performances were evaluated by three evaluators. We analysed the differences in the reliability, objectivity and time efficiency of three different evaluation methods. The calculation of Cronbach's alpha coefficients and analysis of the variance has been used. The analysis of reliability revealed that the combined evaluation method (alpha: 0.928) was the most reliable, the holistic method was less reliable (alpha: 0.879), and the least reliable was the analytical evaluation method (alpha: 0.833). The analysis of objectivity showed that the analytical evaluation method was the most objective (alpha: 0.854), the combined method was less objective (alpha: 0.791), and that the holistic evaluation methods were the least objective method (alpha: 0.778). The analysis of time efficiency revealed that the least time had been spent in the holistic evaluation method and much more in analytical and combined evaluation method. Analysis of the three evaluators, using three different evaluation methods showed no statistically significant differences (p=0.638) between the holistic and combined evaluation method. In contrast, statistically significant differences have been found between the holistic and analytical method (p=0.000) and combined and analytical method (p=0.000). According to the analysis, we can conclude that all three evaluation methods are appropriate for the evaluation of pupils' knowledge in physical education.

Keywords: *Physical education, gymnastics, test task, evaluation guidelines.*

INTRODUCTION

For several years, knowledge evaluation has been among the most influential and simultaneously most complex conceptual educational challenges (Colby & Witt, 2000; Hay, 2006; López-Pastor et al., 2013). Recently, experts have started to emphasize the importance of formative assessment of pupils in which the aim of evaluation is providing the subjects with qualitative feedback information about their knowledge, whilst simultaneously treating each pupil as a unique individual (Blanchard, 2009; Holcar, 2014; Georgakis, Wilson, & Evans, 2015; Leirhaug & Annerstedt, 2015).

Formative assessment is a demanding process, as the primary school teacher evaluates the knowledge of pupils according

to the standards set in the curriculum. Technically speaking, in order to complete the lessons, the task of a teacher is to plan a learning process and teach selected contents in such a way that pupils can acquire them. Additionally, when seen from the aspect of modern paradigm, where pupils are a focal centrepiece, evaluation is merely a means or a guide to their knowledge. In fulfilling this task, teachers have to consider the individual particularities of each pupil. As a result, the teaching process is differentiated and individualized, aiming for pupils to recognize, understand and acquire long-lasting knowledge. This is particularly important in physical education (PE), as the acquired knowledge represents the motor literacy of pupils and as such will undoubtedly influence their sporting participation in their free time and later stages of life (Kovač, Jurak, & Strel, 2003).

In Slovenia, the physical education curriculum (Kovač, & Novak, 2006) sets the guidelines for the monitoring, evaluation and marking of pupils. Monitoring is carried out by guiding pupils from general into specific and in-depth knowledge. Evaluation is carried out by providing feedback to pupils and enabling them to develop and broaden their knowledge, whilst upgrading their general knowledge into more specific knowledge. Marking is carried out through formal appraisal of pupils' knowledge in a form of a score. Marks given should represent an encouragement for pupils to acquire further knowledge. The curriculum sets the standards for the evaluation and marking of learning goals, prescribed at the end of every three-year period. Teachers decide on the goals and contents themselves, as the curriculum provides merely a general framework, which is adjusted to the specific features of the school and lesson realization. In marking, teachers possess wide autonomy, leaving the choice of criteria to their reasoning (Kovač, & Novak, 2006; Plevnik, 2008). Teachers record the marking criteria in the annual working plan and present them to the pupils at the beginning of each academic year. One of the prescribed standards at the end of the second three-year period of primary school is also the knowledge of straddle vault over the buck, performed in a way as described in the present article. The ability of teachers to use different ways of evaluation (holistic, pondered and analytical) facilitates better precision and more objective marking of pupils' knowledge.

In physical education, motor abilities, skills and knowledge are strongly interlinked. Curriculum (Kovač, & Novak, 2006) defines motor abilities as hereditary and being responsible for the execution of movement (e.g. strength, speed, coordination, etc.). In contrast, motor abilities should not be mistaken for acquired skills and learned knowledge (e.g. standing long jump, vault over the buck, etc.).

Teachers agree that in order to evaluate the knowledge of pupils, the most appropriate way is by setting them a task, which will reflect the knowledge acquired according to certain sets standards in the curriculum. In comparison to other academic subjects, PE possesses numerous specific features, as the evaluation comprises both theoretical and practical knowledge as well as the motor abilities of pupils. Assessment of "theoretical" knowledge in PE is done in a conventional fashion consistent with other more established subjects, i.e. by examination, essay or multiple choice questions. Assessment of "practical work" is less easily done. Various practices have emerged, including the use of motor skills and fitness tests, tables of points awarded for performance in areas such as games, swimming and athletics, and the subjective assessment of teachers on matters such as game performance. Assessment of motor skills is mostly done with fitness tests (López-Pastor, 1999, 2006).

Such specifics pose a problem for setting the criteria for evaluation. According to the subject and problem in question, a focus of the study was the evaluation of motor skills in the task of straddle vault over the buck. The summary of various sources on evaluation in similar tasks (Bajec et al., 2002; Dežman & Kovač, 2002; Kovač,

2012; Kovač et al., 2002; Lorenci et al., 2002; Majerič, 2004; Premlč, 2002; Štemberger, 2003; Voglar & Kovač, 2002; Zadražnik, 2002) revealed that teachers most often set the criteria in an analytical and holistic way when making assessments.

The analyses show that in PE many teachers use so-called holistic evaluation for gymnastics, dance, and game performance (Brau-Antony & David, 2002; Estrabaud, Marigneux, & Tixier-Viricel, 2000; Lockwood & Newton, 2004; Kovač, 2012; Majerič, 2004). Teachers assess pupils' skills through observation, using their own professional expertise. The task is evaluated as an entity and is not divided into separate parts. This type of subjective assessment is undoubtedly time-efficient; however, it has several limitations, as it is usually intuitive and adjusted to the level of knowledge and social relationships of the group (Brau-Antony & David, 2002; Estrabaud, Marigneux, & Tixier-Viricel, 2000; Rutar Ilc, 2003). Professional recommendations suggest analytical assessment with the use of evaluation criteria (Newton & Bowler, 2010). For each evaluated task, teachers set precise criteria and descriptions for various parts of it. Nevertheless, some teachers are of the opinion that certain contents cannot be objectively assessed in either a holistic or analytical way (e.g. athletics); therefore, they use a special so-called combined assessment, which includes characteristics of both holistic and analytical types of evaluation (Majerič, 2004; Tomažin et al., 2001a, b, c; 2002). When using this "pondered" type of assessment, teachers consider some parts of the task to be of hierarchical value according to their role in the task. Criteria and descriptions are defined as ponders, ensuring the hierarchical structure according to the importance of each task.

It is also important for teachers to be efficient with the time of evaluation, as the administrative part should not burden them or require too much time from the teaching process. The time should namely be used for the strengthening and expanding of pupils' knowledge. It is estimated that the structured use of all three evaluation methods (holistic, combined and analytical) could also result in better time efficiency and higher quality of lesson realization.

This study has examined an evaluation of one of the most common gymnastics skills: straddle vault over the buck. Numerous authors whose research deals with the assessment of skills in PE agree that the performance of pupils needs to be evaluated with deliberation and by using diverse methods (Ávalos Ramos, Martínez Ruiz, & Merma Molina, 2014; Brau-Antony & David, 2002; Burton, 1998; Kovač, Strel, & Majerič, 2008; Newton & Bowler, 2010).

The main goal of the study was to analyse the differences between three different evaluation methods of the straddle vault over the buck in order to determine the most appropriate way for assessing primary school pupils. Therefore, the measurement characteristics of three different methods (holistic, combined, and analytical) of task evaluation were analysed. As the evaluation is only a part of the systematic teaching process, it should not take too much of the teacher's time; therefore, the time efficiency of each evaluation method has also been observed.

METHODS

The study included 222 boys enrolled in the seventh grade from 11 different Slovenian primary schools, aged 13 years (± 6 months), not exempted from PE classes due to health reasons, and whose parents had given written consent for participation in the research "The Analysis of Children's Development in Slovenia" (Strel et al., 2007). The test sample included 193 boys, whose video recordings were of sufficient quality for the evaluators to be able to assess both attempts.

The gymnastic test task was prepared by Kovač and Čuk (2003) for the purpose of external assessment of PE in the Slovenian school system and transformed for the purpose of this study by Majerič (2004). It included a) descriptions of technically appropriate movement in separate phases of

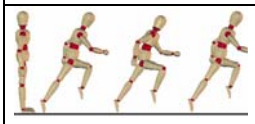
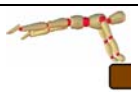
the task and b) criteria with a measurement scale and description of standards. The holistic, combined and analytical evaluation methods were used. A six-level measurement scale (0 to 5 points) was used in all three evaluation methods.

Space: Gymnasium, 18 meters × 2 meters.

Accessories: vaulting buck, 110 cm high; three mats (2 meters × 1 meter), 6 to 12 cm high; springboard, 20 to 25 cm high, 120 cm long, 60 cm wide.

Set-up: run-up distance was optional, allowing the pupils at least 12 meters from the start of the run-up to the springboard, which was placed in front of the vaulting buck at an appropriate distance, by the pupils themselves. Several mats with 6- to 12-cm thickness were placed behind the buck next to each other. The length of the mats was parallel to the axis of the apparatus and at the right angle to the run-up. A third mat was placed on top of these two mats.

Description (technique) and figure of skill straddle vault over the buck Kovač & Čuk (2003)

1) Run-up	2) Hurdle step onto the springboard	3) Take-off from the springboard	4) First flight phase	5) Approach and push-off from the vaulting buck	6) Second flight phase	7) Landing
						
Run-up length is 12 to 14 m long.	Take-off from one foot is followed by landing with two feet on the highest part of a springboard, which is marked with a line. Toes touch the line; arms are behind the body.	Take-off from both feet with arms swinging forward.	The body is extended or slightly piked. Legs straddle just before the contact of hands with the apparatus.	Hands are placed shoulder width, hips travel high above the apparatus, followed by a strong push from the hands.	The upper part of the body is elevated; legs are straddled and placed together just before the landing.	Landing in a still stand with feet together.

Description of evaluation methods

Description of the holistic evaluation method

In the holistic evaluation method, a evaluator assessed the entire presentation of a pupil without "deliberating" the (lack of) knowledge, but merely by "intuitively" forming a mark on the basis of his/her own evaluation standards (Jurman, 1989). For various marks or grades, the teacher simultaneously combined several criteria, which were more realistic according to Rutar Ilc (2003). Criteria were based on the comparison of the quality of a pupil's performance with the ideal (technically correct) model.

Each evaluator marked the first and second attempts. In each attempt, the execution werw classified according to the deviation from correct execution on a scale of 0 to 5. The correct execution were marked with the highest number of points (5) whereas no points (0) were awarded when a pupil did not perform a task or else the execution was not in accordance with the description of individual parts of the task. Attempts that deviated from correct performance were marked on a scale of 1 to 4 according to the description of the task

and the expert knowledge and experience of the evaluator.

Criteria for the holistic evaluation method by Kovač & Čuk (2003)

Criteria for marking:	*Points*	Description of standards
	5	Correct execution.
	4	*Deviation from the correct execution.*
	3	*Deviation from the correct execution.*
	2	*Deviation from the correct execution.*
	1	Incorrect execution.
	0	Not executed or executed not according to the guidelines.
Marking scale:	Mark	Number of points
	1 (unsatisfactory)	0 and 1
	2 (satisfactory)	2
	3 (good)	3
	4 (very good)	4
	5 (excellent)	5

Description notes for a holistic evaluation method

The evaluator assesseed the first and second attempts at the task and noted the number of points in an appropriate field on a form. The attempt with the higher number of points was selected, transformed into a mark, and written in a final mark field. If it was necessary, the evaluator's notes was written in the appropriate field.

Pupil no.:	Attempt	Evaluator's notes	Number of points	Final mark
	1.			
	2.			

Description of the combined evaluation method

In the combined (pondered) method of assessment, a teacher evaluated pupils' attempts of the task whilst considering individual parts of the task having a hierarchical value according to their importance for the execution of the entire task. In this way, the description notes included "ponders", which were defined in a way of ensuring the hierarchical structure according to the importance for the technically correct execution of the movement in the individual task.

In this method, the evaluator assessed both attempts with points from 0 to 10. The movement in each attempt was divided into several phases, each of them with a set maximum number of available points. The evaluator assessed individual phases of movement according to the expert knowledge following the criteria, where 0 represented incorrect execution, and the highest value (ponder) in a certain phase of movement a correct execution was noted with 1, 2 or 3). The total number of points for individual phases of movement were marked down.

Criteria for the combined evaluation method adjusted to Tomažin et al. (2001a, b, c).

Criteria for evaluation of various ponders:	Value of ponder	Description of standards			
		Incorrect execution	Correct execution		
	1	0	1		
	2	0	1	2	
	3	0	1	2	3
Marking scale:	Mark	Sum of ponders (number of points achieved)			
	1 (unsatisfactory)	<3			
	2 (satisfactory)	3-4			
	3 (good)	5-6			
	4 (very good)	7-8			
	5 (excellent)	9-10			
	Final mark				

Description notes for the combined evaluation method adjusted to Tomažin et al. (2001a, b, c).

The evaluator marked points for individual phases of movement and entered the sum: the final score for each attempt onto an evaluation form. The attempt with the higher number of points was transformed into a final mark.

Pupil no.:									
Phase of movement	Highest number of available points in final score (ponder)	Attempt 1				Attempt 2			
		Incorrect execution		Correct execution		Incorrect execution		Correct execution	
1) Rhythmically coordinated run-up	-	not evaluated				not evaluated			
2) Step onto and a take-off from the springboard	2	0	1	2		0	1	2	
3) First flight phase	1	0		1		0		1	
4) Contact and take-off from apparatus	3	0	1	2	3	0	1	2	3
5) Second flight phase	2	0	1	2		0	1	2	
6) Landing	2	0	1	2		0	1	2	
TOTAL NUMBER OF POINTS									
Final mark		1 (unsatisfactory)	2 (satisfactory)	3 (good)	4 (very good)	5 (excellent)			
		<3 points	3-4 points	5-6 points	7-8 points	9-10 points			

Description of the analytical evaluation method

In the analytical method, the teacher first weighed knowledge and a lack of it and finally set a mark according to the evaluation scale and descriptions. Criteria were very precise (multi-level with a description for every level). Such an analytical approach was based on a highly precise identification of deviations (mistakes) from correct execution.

The evaluator assessed both attempts with points from 0 to 5 according to the description notes. In each attempt, the movement was divided into individual phases. Each phase contained the previously defined most common mistakes (see the column "Mistakes" in the execution of movement), which could occur in this part of the task. The evaluator assesseed individual phases according to the table provided. According to their structure, mistakes were divided in technical (deviation of technique from ideal execution) and aesthetic (deviation in elegance and poise of execution) mistakes; according to the severity of deviation, mistakes were either small or large. Small mistakes do not have a significant effect on the execution skill, but rather create a small instability in execution. Large mistakes that significantly influenced the correct execution or else prevent the pupil from performing a skill.

Criteria for analytical evaluation method by Kovač & Čuk (2003)

Measurement scale (points)	Mistakes	Description of standards
5	No or small technical or aesthetic mistakes	Autonomous and reliable execution of straddle vault over the buck without mistakes or with one small technical or aesthetic mistake.
4	Several small technical or aesthetic mistakes	Autonomous and reliable execution of straddle vault over the buck with several small technical or aesthetic mistakes.
3	Several small technical and aesthetic mistakes; one large technical or aesthetic mistakes and several small technical or aesthetic mistakes	Autonomous but not entirely reliable execution of straddle vault over the buck with several small technical and aesthetic mistakes or one large technical and several small technical or aesthetic mistakes.
2	Large technical and/or aesthetic mistakes	Autonomous but not entirely reliable execution of straddle vault over the buck with large technical and/or aesthetic mistakes.
1	Large and small technical and aesthetic mistakes	Execution of straddle vault over the buck in easier circumstances and adjusted way (sit on the buck; help of the teacher needed; fall in transition over the apparatus or landing) with large and small technical and aesthetic mistakes.
0	Not executed or not according to the instructions	Straddle vault over the buck is not performed (run by the buck, stopping in front of the buck).

Mistakes in execution of straddle vault over the buck by Kovač & Čuk (2003)			
1) Run-up		***5) Contact and push-off from the vaulting buck***	
-	Run-up is not being assessed.	TML	Low hips in transition over the buck.
2) Step on the springboard		TML	Very bent arms.
TML	Step onto the first part of springboard.	TML	Push-off is too late and weak.
TML	Step on the springboard with body leaning forward too much.	TML	Insufficient straddling of the legs and touching of buck with legs.
TML	Step on with very bent legs.	TMS	Insufficiently elevated hips.
TML	Take-off from flat feet.	TMS	Slightly bent arms.
TMS	Step on the last but not optimum part of springboard.	TMS	Hands are not parallel on the buck.
TMS	Wrong arm movement (arms upwards before the take-off).	AML	Very bent and completely relaxed legs and feet.
AML	Completely relaxed body.	AMS	Slightly bent and relaxed legs and feet.
AMS	Slightly relaxed body.	N	Fall off the buck.
N	Run by the springboard.	***6) Second flight phase***	
3) Take-off from springboard		TML	Short and low flight.
TML	Very poor execution of take-off (very bent legs in the air after take-off).	TMS	Insufficient straddle of the legs.
TML	Low take-off.	TMS	Swing with straddled legs forward.
TML	Hesitation on the springboard before take-off.	TMS	Body is not upright before landing.
TMS	Take-off is not completed (legs are slightly bent in the air after take-off).	AML	Very bent and completely relaxed legs and feet.
AML	Completely relaxed body, very bent legs and feet.	AMS	Slightly bent and relaxed legs and feet.
AMS	Slightly relaxed body, very bent legs and feet.	***7) Landing***	
N	Run and stop at the vaulting buck.	TML	Loss of balance: two or more additional steps.
4) First flight phase		TML	Landing in a deep squat.
TML	Springboard is too close to the buck, and the flight is low.	TMS	Landing with legs straight (non-elastic).
TML	Too early contact with hands on the apparatus.	TMS	Loss of balance: additional step.
TMS	Incorrect arm swing (too high or too low).	TMS	Landing outside of direction.
TMS	Too early straddling of the legs in flight.	AML	Completely relaxed body.
AML	Very bent and completely relaxed legs and feet.	AMS	Slightly bent arms; relaxed head position; relaxed body.
AMS	Slightly bent and relaxed legs and feet.	N	Fall at landing.
TMS – Technical mistake (small); TML – Technical mistake (large); AMS – Aesthetic mistake (small); AML – Aesthetic mistake (large).			

Description notes for analytical evaluation method

The evaluator totalled the mistakes and entered them in the field "sum of mistakes" for the first and second attempts, separately. The better attempt (with fewer mistakes) was marked, and points were transformed into the final mark.

Pupil no.:	Mistakes		Sum of mistakes		Final mark
			Attempt 1	Attempt 2	
	TML				
	TMS				
	AMS				
	AML				

TMS – Technical mistake (small): TML – Technical mistake (large); AMS – Aesthetic mistake (small); AML – Aesthetic mistake (large).

After warming up, six different test tasks (two gymnastics, two track and field, one basketball, and one volleyball) were explained and demonstrated to pupils who then performed each test task three times under the same conditions. The second and third attempts were video recorded; the study examines the evaluation of the gymnastics task straddle vault over the buck.

Three PE teachers evaluated the performance of tasks with a use of each protocol. For the purpose of correct evaluation, they received special training. Before the assessment, evaluators carefully read the description and the evaluation criteria of the task. Afterward, they independently evaluated both video-recorded performances in normal speed. The recordings could not be stopped, and the evaluators could not view them in slow motion or more than once. For evaluation, three different evaluation methods were used. First, the performances were evaluated with the holistic method, then with the combined method, and finally with analytical evaluation method. For each pupil, the better score of two attempts was used for further statistical analysis.

In order to monitor the reliability of the study, all three evaluators repeated the evaluation of the first ten performances at 30-day intervals. In order to examine the time efficiency of the different evaluation methods, the time needed for evaluation has been measured three times. First, time measuring was carried out for the first twenty measured subjects, then for twenty measured subjects from the middle of the sample (subjects 100–120) and finally for the last twenty subjects from the sample (subjects 203–222).

Data were processed with the use of SPSS for Windows software. Cronbach's reliability coefficient alpha and calculation of the concordance between respective evaluator's grades and the common test object were used for the evaluation of reliability and objectivity. Analysis of the variance was used to analyse the differences in scores between three evaluators. All statistics used an alpha level of $p < 0.05$.

RESULTS

Reliability of evaluation

Table 1
Reliability of evaluation – descriptive (simple) statistic parameters.

		Simple statistics – first evaluation						Simple statistics – control evaluation				
		N	min	max	M	SD		N	min	max	M	SD
HMA	E 1	10	3.00	5.00	4.00	0.47	CE1	10	3.00	4.00	3.30	0.48
	E 2	10	2.00	4.00	3.20	0.79	CE 2	10	2.00	4.00	3.10	0.88
	E 3	10	2.00	4.00	3.10	0.74	CE 3	10	2.00	4.00	3.10	0.88
	AOE	10	3.00	4.33	3.43	0.47	CA AOE	10	2.33	4.00	3.16	0.57
CMA	E 1	10	2.00	4.00	3.70	0.48	CE1	10	3.00	4.00	3.40	0.52
	E 2	10	1.00	4.00	3.00	0.82	CE 2	10	2.00	4.00	2.90	0.74
	E 3	10	3.00	5.00	3.90	0.57	CE 3	10	2.00	4.00	3.20	0.63
	AOE	10	2.67	4.00	3.53	0.45	CA AOE	10	2.33	4.00	3.16	0.50
AMA	E 1	10	2.00	3.00	2.40	0.52	CE1	10	2.00	4.00	2.50	0.71
	E 2	10	2.00	4.00	2.70	0.82	CE 2	10	2.00	4.00	2.70	0.82
	E 3	10	2.00	5.00	3.60	0.84	CE 3	10	2.00	4.00	3.30	0.82
	AOE	10	2.00	4.00	2.90	0.62	CA AOE	10	2.33	3.67	2.83	0.52

Key: HMA – holistic evaluation method, CMA – combined evaluation method, AMA – analytical evaluation method, E 1 – first evaluator, E 2 – second evaluator, E 3 – third evaluator; AOE – average of three evaluators; CA E1 - first evaluator - control (second) evaluation; CA E2 – second evaluator - control (second) evaluation; CA E3 – third evaluator - control (second) evaluation; CA AOE – average of three evaluators - control (second) evaluation.

Table 2
Reliability of evaluation– correlations between the evaluators.

			Correlation between the marks									
HMO	E 1	E1	1.000		E 2	1.000		E 3	1.000		AOE	1.000	
	E 2	CA E1	0.488	1.000	CA E2	0.933	1.000	CA E3	0.498	1.000	CA AOE	0.799	1.000
	AOE	alpha	0.655		alpha	0.962		alpha	0.659		alpha	0.879	
CMO	E 1	E 1	1.000		E2	1.000		E 3	1.000		AOE	1.000	
	E 2	CA E1	0.534	1.000	CA E2	0.922	1.000	CA E3	0.371	1.000	CA AOE	0.872	1.000
	AOE	alpha	0.695		alpha	0.956		alpha	0.539		alpha	0.928	
AMA	E 1	E 1	1.000		E 2	1.000		E 3	1.000		AOE	1.000	
	E 2	CA E1	0.304	1.000	CA E2	1.000	1.000	CA E3	0.352	1.000	CA AOE	0.725	1.000
	AOE	alpha	0.449		alpha	1.000		alpha	0.520		alpha	0.833	

Key: HMA – holistic evaluation method, CMA – combined evaluation method, AMA – analytical evaluation method, E 1 – first evaluator, E 2 – second evaluator, E 3 – third evaluator; AOE – average of three evaluators; CA E1 - first evaluator - control (second) evaluation; CA E2 – second evaluator - control (second) evaluation; CA E3 – third evaluator - control (second) evaluation; CA AOE – average of three evaluators - control (second) evaluation.

Table 3
Objectivity of evaluation – descriptive (simple) statistic parameters.

Evaluation method	Evaluator	min	Max	M	SE	SD
HMA	E 1	1.00	5.00	3.22	1.02	7.34
	E 2	1.00	5.00	2.12	1.10	7.90
	E 3	1.00	5.00	3.29	1.10	7.94
	SUM	1.00	5.00	2.88	1.20	4.98
CMA	E 1	1.00	5.00	3.18	0.90	6.46
	E 2	1.00	4.00	1.83	0.83	5.94
	E 3	1.00	5.00	3.74	0.97	6.99
	SUM	1.00	5.00	2.92	1.20	5.00
AMA	E 1	1.00	4.00	1.78	0.80	5.78
	E 2	1.00	4.00	1.77	0.83	6.00
	E 3	1.00	5.00	2.57	1.04	7.52
	SUM	1.00	5.00	2.04	0.97	4.05

Key: HMA – holistic evaluation method, CMA – combined evaluation method, AMA – analytical evaluation method, E 1 – first evaluator, E 2 – second evaluator, E 3 – third evaluator; AOE – average of three evaluators

Table 4
Objectivity of evaluation – correlation between evaluators.

		Correlation between evaluators and between the evaluators and K1				Communalities
HMA	E 1	1.000	0.613	0.586	0.888	0.788
	E 2	0.613	1.000	0.430	0.815	0.664
	E 3	0.586	0.430	1.000	0.799	0.638
	K1	λ	cum %		alpha	
		2.090	69.680		0.778	
CMA	E 1	1.000	0.639	0.568	0.880	0.774
	E 2	0.639	1.000	0.486	0.842	0.710
	E 3	0.568	0.486	1.000	0.804	0.647
	K1	λ	cum %		alpha	
		2.131	71.021		0.791	
AMA	E 1	1.000	0.834	0.559	0.903	0.815
	E 2	0.834	1.000	0.667	0.943	0.889
	E 3	0.559	0.667	1.000	0.822	0.676
	K1	λ	cum %		alpha	
		2.38	79.323		0.854	

Table 5
Time efficiency of evaluation

	HMA				CMA				AMA			
Time of evaluation	E 1	E 2	E 3	AOE	E 1	E 2	E 3	AOE	E 1	E 2	E 3	AOE
subjects 1 to 20	9.10	17.55	10.30	12.32	14.12	26.13	23.53	21.26	17.14	25.18	19.38	20.57
subjects 100 to 120	8.18	9.15	8.30	8.54	12.26	14.38	20.13	15.59	10.42	19.25	9.05	12.91
subjects 203 to 222	8.17	8.10	7.10	7.79	11.29	17.19	18.32	15.60	11.29	16.49	8.47	12.08

Key: HMA – holistic evaluation method, CMA – combined evaluation method, AMA – analytical evaluation method; E 1 – first evaluator, E 2 – second evaluator, E 3 – third evaluator; AOE – average of three evaluators; time of evaluation in minutes

Table 6
Analysis of the variance between the holistic, pondered and analytical evaluation methods.

Comparison	Type of evaluation	Simple statistics					Homogeneity variance test		Analysis of the variance	
		N	Min	max	M	SD	F	Sig.	F	Sig.
HMA/	HMA	193	1.00	5.00	2.88	0.89				
AMA	AMA	193	1.00	4.33	2.04	0.79				
	Total	386	1.00	5.00	2.46	0.94	2.340	0.127	95.328	0.000
HMA/	HMA	193	1.00	5.00	2.88	0.89				
CMA	CMA	193	1.00	4.67	2.92	0.76				
	Total	386	1.00	5.00	2.90	0.83	5.791	0.017	0.222	0.638
CMA/	CMA	193	1.00	4.67	2.92	0.76				
AMA	AMA	193	1.00	4.33	2.04	0.79				
	Total	386	1.00	4.67	2.48	0.89	0.891	0.346	124.402	0.000

Key: HMA – holistic evaluation method, CMA – combined evaluation method, AMA – analytical evaluation method

Key: HMA – holistic evaluation method, CMA – combined evaluation method, AMA – analytical evaluation method, E 1 – first evaluator, E 2 – second evaluator, E 3 – third evaluator Analysis (Table 1) of the mean values of the average marks of three evaluators revealed the highest marks in the first evaluation by all three methods (HMA mean of marks in first evaluation: 3.43, mean of marks in control evaluation: 3.16; CMA mean of marks in first evaluation: 3.53, mean of marks in control evaluation: 3.16; AMA mean of marks in first evaluation: 2.90, mean of marks in control evaluation: 2.83) This has shown that all three evaluators gave lower marks in the second (control) evaluation. The analysis (Table 2) revealed that the combined evaluation method was the most reliable since Cronbach's alpha coefficient was greatest for the average mark of the three evaluators (alpha: 0.928); the holistic method was less reliable (alpha: 0.879), whilst the least reliable method was the analytical evaluation method (alpha: 0.833). Nevertheless, the Cronbach's alpha coefficient was high enough that the reliability of the evaluators was adequate in all three evaluation methods could be observed.

Objectivity of evaluation

The analysis (Table 3) showed that the third evaluator (mean values of the average marks by HMA: 3.29, CMA: 3.74, AMA: 2.57) gave the highest marks, the first a little lower (mean values of the average marks by HMA: 3.22, CMA: 3.18, AMA: 1.78) and the second the lowest (mean values of the average marks by HMA: 2.12, CMA: 1.83, AMA: 1.77). This has been typical and evident with all three evaluation methods. Compatibility between the scores of individual evaluators and the common object of assessment (the first main component of the scores of all three evaluators) was verified in order to monitor the objectivity of the evaluation. The analysis (Table 4) revealed that the analytical evaluation method was the most objective because Cronbach's alpha coefficient was the highest (alpha: 0.854); the combined method was less objective (alpha: 0.791), whilst the least objective method was the holistic evaluation methods (alpha: 0.778). Nevertheless, the Cronbach's alpha coefficient was high enough that it could be observed that the objectivity of the evaluators was adequate in all three evaluation methods.

Time efficiency of evaluation

The average evaluation time of all three evaluators revealed that the least time had been spent in the holistic evaluation method (subject 1 to 20 evaluation: 9.10 minutes, subject 100 to 120: 8.18 minutes, subject 203 to 222: 8.17 minutes). The most time has been spent in the combined evaluation method (subject 1 to 20 evaluation: 21.26

minutes, subject 100 to 120: 15.59 minutes, subject 203 to 222: 15.60 minutes). It is interesting that the combined evaluation method took more time for evaluation than the analytical method did (subject 1 to 20 evaluation: 20.57 minutes, subject 100 to 120: 12.91 minutes, subject 203 to 222: 12.08 minutes), which has been the most complex (Table 5). The analysis showed that the evaluations with all three methods were economical. Evaluators for the knowledge evaluation of 20 pupils took less than half of the school lesson time, which was defined as 45 minutes.

Analysis of differences between the holistic, pondered, and analytical evaluation methods of motor skills

The arithmetical mean value (evaluation marks) was the highest with the combined evaluation method (2.92), followed by the holistic (2.88) and analytical methods (2.04), indicating the analytical method to be the most critical. Differences in average marks between the three evaluators, using the holistic, combined, and analytical evaluation methods, have been examined with the use of analysis of the variance. No statistically significant differences (p=0.638) have been found between the marks acquired with the holistic and combined evaluation methods (see Table 6). In contrast, statistically significant differences have been found between the holistic and analytical marks (p=0.000) as well as between the combined and analytical marks (p=0.000).

DISCUSSION

School marks would be reliable if the same teachers in the re-evaluation of knowledge would give for equal marks the same knowledge (result) (Marentič Požarnik, 2000). To analyse the reliability of the evaluation in our test, all three evaluators assessed the first ten performances of the straddle vault over the buck (out of 222) twice within a 30-day interval. It has been observed that the reliability of the evaluators was adequate in all evaluation methods. The analysis (see Table 2) revealed that the combined evaluation method was the most reliable since Cronbach's alpha coefficient was greatest for the average mark of the three evaluators (0.928); the holistic method was less reliable (0.879), whilst the least reliable method was the analytical evaluation method (0.833). We have found that in our case the reliability of the evaluation was very good. For the determination of the reliability of the measurement of constructed variables the criterion by Ferligoj, Leskovšek and Kogovšek was normally used (1995). This criterion indicated the reliability of the measurement as very good, if the Cronbach's alpha coefficient was greater than or equal to 0.80, and as good if it was greater than or equal to 0.70. It has also been revealed that all three evaluators were more critical in the second (repeated) evaluation (see Table 1), indicating better insight into the perception of mistakes due to greater experience with evaluation. The extremely high reliability of the second evaluator has been observed in all three evaluation methods, in particular in the analytical method. Significantly, the lower reliability of the other two evaluators was evident, particularly in the analytical method, confirming the findings of Marentič Požarnik (2000), who stated that evaluation with the analytical method is very reliable when the criteria are well known, whilst decreases in the criteria awareness and experience also result in reduced reliability. The findings of this study are interesting, as they indicate that the holistic evaluation method is relatively reliable despite having the least defined evaluation criteria. These confirm findings regarding the measuring characteristics (Brau-Antony & David, 2002; Estrabaud, Marigneux, & Tixier-Viricel, 2000; Lockwood & Newton, 2004; Majerič, 2004), in which the authors recommended the holistic evaluation method in gymnastics. Other researchers (Bajec et al., 2002; Dežman & Kovač, 2002; Kovač, 2012; Kovač et al., 2002; Lorenci et al., 2002; Majerič, 2004; Premlč, 2002; Štemberger, 2003; Voglar & Kovač, 2002;

Zadražnik, 2002) who analysed the reliability of analytical and holistic evaluation also reported that such evaluations were reliable enough for school assessment. The values of Cronbach's alpha coefficients in these studies were according to Ferligoj et. al (1995) comparable with our results, nevertheless some researchers calculated in different test task slightly higher values of Cronbach's alpha coefficients 0.950 (Bajec et. al, 2002); 0.987 (Dežman & Kovač, 2002), 0.980 (Kovač et al., 2002); 0.970 (Lorenci et al, 2002); 0.994 (Zadražnik, 2002). We assumed that the differences in slightly higher values of Cronbach's alpha coefficients in the cited studies were due to different times that had elapsed between the first and second evaluations (a few days in these studies, 30 days in our study); different number of evaluators, motivation and special skills and expertise of evaluators (in these studies, the tasks were evaluated by specialists for individual sports; in our study physical education teachers from practice). From this point of view, our study was closer to real school practice.

The evaluation would be objective if the pupils got the same marks for the same results by different evaluators. Jurman (1989) cited various studies and concluded that Cronbach's alpha coefficients between different evaluators were between 0.70 to 0.80. He marked this values as good. In our case, regarding the objectivity of evaluation for the average evaluators' marks, the results revealed (see Table 4) the highest Cronbach's alpha coefficient in the analytical evaluation method (0.854), followed by the combined (0.791) and holistic evaluation method (0.778). Similar consistency between the evaluators was also observed by Majerič, Kovač, Dežman and Strel (2005) in an evaluation of long jump (holistic evaluation method Cronbach's alpha: 0.809; combined evaluation method Cronbach's alpha: 0.811; analytical evaluation method Cronbach's alpha: 0.836). Quite similar results have (Cronbach's alpha: 0.880) been found by Bajec, Bon, Dežman and Kovač (2002) in the analytical evaluation method test of throwing the ball to the goal in handball. Therefore, it can be concluded that with better-defined criteria PE teachers could more objectively evaluate different motor skills. From the perspective of the formative monitoring of pupils, this information is important because specifically defined criteria provide precise feedback information about their knowledge whilst simultaneously signalling to the other subjects that the mark is objective for all and thus unbiased. The subjectivity of the teacher could, therefore, be largely excluded.

In determining the time efficiency of evaluation, we considered the time spent by three different evaluators while they evaluated the pupil's knowledge by three different evaluation methods. The average evaluation time of all three evaluators revealed that the least time had been spent in the holistic method and the most time in the combined evaluation method (see Table 5). Such results were expected, as the criteria were the simplest in the holistic and most complex in the combined and analytical evaluation method. In comparison to the first timed period (subjects 1 to 20), the evaluation time to the last timed period (the last twenty subjects) was reduced in all three evaluation methods. Specifically, in the holistic and analytical evaluation methods, the time was nearly halved, whereas it was shortened by a third in the pondered (combined) method. It can be concluded that in the first batch of measured subjects (1 to 10), individual evaluators were acquiring evaluation skills for the set criteria, whereas in the second (100–120) and third (203–222) timed batches of subjects' work had already been carried out routinely. According to the data and theoretical suggestions of several authors (Airasian, 1996; Burton 1998; Rutar Ilc, 2000, 2003), it has been concluded that evaluation is predominantly organizational and thus a technical challenge, which could be carried out with higher time efficiency by providing continuous training for teachers. As the time-efficient evaluation procedures

are those that with the sensible use of time and energy provide the highest quality results (possess good measuring evaluation characteristics) (Marentič Požarnik, 2000), time efficiency also needs to be considered in correlation with the reliability and objectivity of evaluation. Specifically, even the most time-efficient evaluation methods are not justified without reliability and objectivity (Kodelja, 2000). In our case, all three methods were economical. Evaluators for the knowledge evaluation of 20 pupils took less than half of the school lesson time, which was defined by 45 minutes. We did not find similar studies that would identify the time efficiency of evaluation of the test tasks.

Differences in the average marks between the three evaluators, using the holistic, combined, and analytical evaluation methods, were found between the analytical and holistic and between the analytical and combined evaluation method. In an evaluation of long jump, Majerič, Kovač, Dežman and Strel (2005) found differences in all three evaluations. We did not find other similar studies that analyse the differences between holistic, combined, and analytical evaluation methods. In our case, we have found that the arithmetical mean value was the highest with the combined evaluation method, followed by the holistic and analytical methods, indicating the analytical method to be the most critical. When comparing the values expressed in the form of school marks, the difference between the average values is a full mark, which is quite considerable. In simple terms, pupils would receive a mark of 4 (very good) for the evaluated knowledge when the holistic and combined methods are used and only a mark of 3 (good) when evaluated with the analytical method. Consequently, and considering the modern paradigm of the formative monitoring of pupils, the authors recommend that teachers in the teaching process for 13-year-old pupils to use more detailed, i.e. analytical criteria in the evaluation of knowledge. Pupils will, as a result, receive feedback information about their knowledge; they will recognize their mistakes and understand what needs to be improved. In the formal assessment, the authors recommend that teachers to use criteria in the holistic or combined methods whilst still paying conscious attention to adequate objectivity. The low marks received were a result of straddle vault over the buck being quite a demanding skill for 13-year-old pupils, particularly when performed autonomously (without the assistance of the teacher). Pupils have to connect a run-up, takeoff from a springboard from two feet and the contact/push-off with the arms from the apparatus, whereas the second flight phase has to be high and adequately long with legs straddled and straight; all movement has to finish in a stable landing. Due to the progressively lower motor abilities of pupils, which are reflected in decreased muscular strength of arms and shoulders and power strength (Strel et al., 2007), pupils could experience difficulties in the take-off from springboard and consequently with arms from the apparatus. As a result, the flight is low and short, resulting in low marks for the executed task.

CONCLUSION

Gymnastic contents are included in all PE curricula and at each level of education (Živčić Marković, Sporiš, & Čavar, 2011). In recent primary school PE curricula, gymnastics remains one of the most important elements around the world (Hardman, Murphy, Routen, & Tones, 2014), as it offers a great range of locomotive, stability and body control movements, which are highly important for the development of children (Kovač & Novak, 2001; Živčić Marković et al., 2011). Gymnastics requires a great diversity of movements: control of body movement during transitions from dynamic to static elements and vice versa, and body balance during frequent changes of the body position in space (Novak, Kovač, & Čuk, 2008; Živčić Marković et al., 2011).

Jumps and vaults are very important in the development of children. The straddle vault over the buck is one of the most common items in PE contents in all grades. Bučar et al. (2010) reported that more than 94% of PE teachers implemented this vault in the last three grades of Slovenian primary school. By including different vaults in the lessons plan, teachers will be able to improve or, at a minumum, maintain the level of motor abilities in their pupils throughout the years. Successful performance of vaults requires accurate muscular activity of specific intensity (muscular strength in arms and shoulders; explosive strength of legs during the take-off from springboard), the right moment (timing) during the take-off from the vault and flexibility (during the flight phase) and balanced landing (Novak et al., 2008).

All three evaluation methods for the straddle vault over the buck showed high reliability and objectivity evaluation, indicating the appropriate selection of test criteria and descriptions. Some differences between the three evaluation methods were not significant. Nevertheless, data for the evaluation of straddle vault over the buck revealed that the pondered evaluation method is the most reliable with regards to the measuring characteristics, whereas the analytical evaluation method is the most objective, and the holistic evaluation method the most time-efficient.

The measuring characteristics of all three evaluation methods were revealed to be appropriate; therefore, in conformity with the autonomy of teachers, it is mostly up to them to decide which evaluation method they will use as long as it is adjusted to the knowledge level of their pupils. For formative assessment, the measuring scales and criteria should be different according to the purpose of evaluation (internal, external), the developmental stage of pupils, and the complexity of the evaluated movement. As a result, the authors recommend that teachers use the analytical or combined methods in the monitoring stage of the teaching process. A prepared analytical or combined model tasks with a description of movement, common mistakes and precise criteria focuses on the learning of each pupil whilst providing suitable feedback. The process can serve as an important function in further teaching, as teachers could identify the problems of pupils and adapt the teaching process. Well-learnt gymnastics skills can generate feelings of satisfaction in pupils and encourage the practice of physical activity (Šimůnková, Novotná, & Chrudimsky, 2013). In the final formal assessment of the skill, teachers should use the most time-efficient, i.e. the holistic evaluation method, for this age group, as it will allow more time for the previous phases in the teaching process.

Teachers give the greatest importance to correct technique in gymnastics skills (Ávalos Ramos et al., 2014); therefore, the task, selected for the evaluation in the present study by three evaluators, also placed an emphasis on the technically correct execution. At the same time, Ávalos Ramos, Martínez Ruiz, and Merma Molina (2014) pointed out great contradictions in the evaluation of school gymnastics. The divergences between the use and evaluation of learning activities indicate that teachers do not employ a great deal of reflection in their planning, nor in their decision-making (Tsui, 2009). As a result, teachers need to be adequately prepared for evaluation, as the process of evaluation itself can be considered a skill (AAHPERD, 1999; Burton, 1998; Pangrazi, 1998) that can be developed in PE teachers. In order to develop this skill, continuous training in various evaluation methods has to be provided at conferences or by using various material (e.g. video recordings on the internet). Undoubtedly, quality teaching is of key importance, as only then the pupils will acquire diverse motor skills necessary for their physical and motor development.

We can confirm that the major weakness of the study was the evaluation process. The evaluators did not evaluate the knowledge of the pupils in real school situations, but the knowledge recorded on videotape. This type of evaluation was

rarely used in practice. However, we found useful information from many teachers. It is important to point out that the evaluators were teachers in our study, while the evaluators were sport experts in other similar studies. From this perspective, it can be concluded, that all three evaluation methods were good tools for teachers to evaluate pupils' knowledge. Despite the reliability, objectivity and time efficiency being found in all three evaluation methods, in the future the assessment procedures in PE will need to be even more adjusted to the spirit of modern formative monitoring of pupils, encouraging the evaluation in very authentic situations. A formative assessment instrument (the assessment wheel) supports a constructivist perspective in which pupils take increasing responsibility for what is learned and how it is represented (MacPhail & Halbert, 2010). An assessment wheel is a simple form of pupil self-assessment, encouraging the pupil to record, reflect on, and map their learning to the rich task and to assess their progress towards a pre-set goal. It also identifies any learning gaps that may exists and enables pupils to plan for the next phase of their learning as well as providing a context of feedback. According to López-Pastor et al. (2013), this also signifies a move away from "test" culture to an "assessment culture" in the new paradigm of "assessment for learning".

REFERENCES

Airasian, P. W. (1996). *Assessment in the Classroom*. New York: New York: McGraw-Hill, Inc.

American Alliance for Health, Physical Education, Recreation and Dance (AAHPERD) (1999). *Physical Education for Lifelong Fitness. The Physical Best Teacher's Guide*. Champaign: Human Kinetics.

Ávalos Ramos, M. A., Martínez Ruiz, M. A., & Merma Molina, G. (2014). Inconsistencies in the curriculum design of educational gymnastics: case study. *Science of Gymnastics Journal 6*(3), 23–37.

Bajec, D., Bon, M., Dežman, B., & Kovač, M. (2002). Ocenjevanje praktičnega znanja v rokometu pri pouku športne vzgoje. [*Evaluation of practical handball skills in physical education lessons*] V R. Pišot, V. Štemberger, F. Krpač, & T. Filipčič (Eds.), *Zbornik 2. mednarodnega znanstvenega in strokovnega posveta Otrok v gibanju* (pp. 170–175). Ljubljana: Pedagoška fakulteta.

Blanchard, J. (2009). *Teaching, Learning and Assessment*. Maidenhead: McGraw-Hill Education.

Brau-Antony, P. S. & David, B. (2002). Les modèles en EPS. *Éducation physique et sport, 53*(297), 79–83.

Bučar Pajek, M., Čuk, I., Kovač, M., & Turšič, B. (2010). Implementation of the gymnastics curriculum in the third cycle of basic school in Slovenia. *Science of Gymnastics Journal, 2*(3), 15–27.

Burton, A. W. (1998). *Movement Skill Assessment.* Champaign: Human Kinetics.

Colby, J. & Witt, M. (2000). *Defining Quality in Education*. A paper presented by UNICEF at the meeting of The International Working Group on Education. Florence, Italy, June 2000. New York, USA: United Nations Children's Fund.

Dežman, B. & Kovač, M. ***(2002). Zanesljivost in objektivnost ocenjevanja praktičnega znanja v košarki pri pouku športne vzgoje.*** [Reliability and objectivity of evaluating practical basketball skills in physical education lessons] V R. Pišot, V. Štemberger, F. Krpač, & T. Filipčič (Eds.), Zbornik 2. mednarodnega znanstvenega in strokovnega posveta Otrok v gibanju (p. **201–206).** Ljubljana: Pedagoška fakulteta.

Estrabaud, P. P., Marigneux, C., & Tixier-Viricel, C. (2000). Baccalauréat un example pratique d' évulation. Éducation physique et sport, 51(284), 23–-25.

Ferligoj, A., Leskovšek, K., & Kogovšek, T. (1995). Zanesljivost in veljavnost merjenja. [**Reliability and validity of measurement**]. Ljubljana: Fakulteta za družbene vede.

Georgakis, S., Wilson, R., & Evans, J. (2015). Authentic Assessment in Physical

Education: A Case Study of Game Sense Pedagogy. *Physical Educator, 72*(1*), 67–86.*

Hardman, K., Murphy, C., Routen, A., & Tones, S. (2014). *World-wide Survey of School physical education. Final Report*. Paris. UNESCO.

Hay, P. G. (2006). *Assessment for Learning in Physical Education. The handbook of PE.* London, Sage.

Holcar, A. H. (2014). Začetki formativnega spremljanja v slovenskem prostoru. [*Beginnings of formative monitoring in Slovenia*] *Vzgoja in izobraževanje (Priloga), 25*(7–8)

Jurman, B. (1989). *Ocenjevanje znanja. Selekcija in orientacija učencev.* [*Evaluation of knowledge. Selection and orientation of pupils.*] Ljubljana: DZS.

Kodelja, Z. (2000). Pravičnost in ocenjevanje. [*Justification and evaluation*] V J. Krek & M. Cencič (Eds.), *Problemi ocenjevanja in devetletna osnovna šola* (pp. 15–23). [*Problems of evaluation in primary school*] Ljubljana: Pedagoška fakulteta Univerze v Ljubljani in Zavod RS za šolstvo.

Kovač, M. & Čuk, I. (2003). *Testna naloga za preverjanje praktičnih znanj pri zunanjem preverjanju in ocenjevanju znanja ob zaključku devetletne osnovne šole: gimnastika – raznožka čez kozo. [Test task for evaluation of practical skills in external evaluation and assessment of knowledge at the end of primary school: gymnastics – straddle vault over the buck]* Ljubljana: Republiški izpitni center.

Kovač, M. & Novak, D. (2001). *Učni načrt za osnovno šolo. Športna vzgoja*. [*Primary school curriculum. Physical education*] Ljubljana: Zavod RS za šolstvo.

Kovač, M. (2012). Assessment of gymnastic skills at physical education – the case of backward roll. *Science of gymnastics journal. 4*(3), 25–35.

Kovač, M., Jurak, G., & Strel, J. (2003). Predlog modela in meril notranjega preverjanja in ocenjevanja znanja pri športni vzgoji. [*A proposal of model and criteria for internal evaluation and assessment in physical education*] *Šport, 51*(2), 21–27.

Kovač, M., Strel, J., & Majerič, M. (2008). Conceptual dimensions of evaluation and assessment in physical education – reasons for using different standards and criteria. In I. Prskalo, V. Findak, & J. Strel (Eds.), *Kinesiological education - answer of the contemporary school* (pp. 7–25). Zagreb: Učiteljski fakultet Sveučilišta u Zagrebu.

Kovač, M. & Novak, D. (2006). *Učni načrt za osnovno šolo [The curriculum for primary school]*. Ljubljana: Urad za šolstvo. Predmetna kurikularna komisija za športno vzgojo.

Kovač, M., Žakelj, M., Čuk, I., Dežman, B., Voglar, M., & Bučar, M. (2002). Ocenjevanje praktičnega znanja gimnastike pri pouku športne vzgoje. [*Assessment of practical gymnastics skills in physical education lessons*] V R. Pišot, V. Štemberger, F. Krpač, & T. Filipčič (Eds.), *Zbornik 2. mednarodnega znanstvenega in strokovnega posveta Otrok v gibanju* (pp. 280–285). Ljubljana: Pedagoška fakulteta.

Leirhaug, P. E. & Annerstedt, C. (2015). Assessing with new eyes? Assessment for learning in Norwegian physical education. *Physical Education and Sport Pedagogy 20*(1–36)

Lockwood, A. & Newton, A. (2004). Assessment in PE. In S. Capel, *Learning to teach Physical Education in the Secondary School*: *A companion to school experience*. 2nd Edition. Abingdon: Routledge.

López-Pastor, V. M. (1999). *Prácticas de Evaluació en Educación Fisica: estudio de casos en primaria secundarioa y formación del profesorado*. Valladolid: Universidad de Valladolid.

López-Pastor, V. M. (2006). *La Evaluación en Educación Fisica: Revisión de los modelos tradicionales y planteamiento de una alternativa: La evaluación formativa y compartida*. Valladolid: Universidad de Valladolid.

López-Pastor, V. M., Kirk, D., Catalan, E. L., MacPhail, A., & Macdonald, D. (2013). Alternative assessment in physical education: a review of international literature. *Sport, Education and Society, 18*(1), 57–76.

Lorenci, B., Kovač, M., & Dežman, B. (2002). Ocenjevanje praktičnega znanja iz atletike pri pouku športne vzgoje. [*Assessment of practical athletics skills in physical education lessons*] V R. Pišot, V. Štemberger, F. Krpač, & T. Filipčič (Eds.), *Zbornik 2. mednarodnega znanstvenega in strokovnega posveta Otrok v gibanju* (pp. 299–304). Ljubljana: Pedagoška fakulteta.

Majerič, M. (2004). Analiza ocenjevanja športnih znanj pri športni vzgoji. [*An analysis of evaluation of sports knowledge in physical education*] *Doktorska disertacija* [Doctoral thesis], Ljubljana: Fakulteta za šport.

Majerič, M., Kovač, M., Dežman, B., & Strel, J. (2005). Analysis of three different ways of assessing motor abilities with the testing assignment of long jump with approach. In D. Milovanović & F. Prot (Eds.), *Proceedings book. 4th International Scientific Conference on Kinesiology* (pp. 98–102). Zagreb: Faculty of Kinesiology, University of Zagreb.

Maretič Požarnik, B. (2000). Ocenjevanje učenja ali ocenjevanje za (uspešno) učenje? Kako zmanjšati neskladje med nameni in učinki ocenjevanja. [*Evaluation of learning or evaluation for (successful) learning. How to reduce imbalance between the purpose and effects of evaluation?*] *Vzgoja in izbraževanje, 31*(2-3), 3–9.

Newton, A. & Bowler, M. (2010). Assessment in PE. In S. Capel & M. Whitehead (Eds.) *Learning to Teach Physical Education in the Secondary School: A Companion to School Experience*. 3rd Edition. London: Routledge.

Novak, D., Kovač, M., & Čuk, I. (2008). *Gimnastična abeceda. [ABC of gymnastics]* Ljubljana: Fakulteta za šport.

Pangrazi, R. P. (1998). *Dynamic Physical Education for Elementary School Children.* (12th edition). Toronto: Allyn and Bacon.

Plevnik, T. (2008). *Ravni avtonomije in odgovornosti učiteljev v Evropi: Eurydice. [*Levels of Autonomy and Responsibilities of Teachers in Europe: Eurydice*]*. Ljubljana: Ministrstvo za šolstvo in šport.

Premlč, M. (2002). *Oblikovanje in izdelava merskih postopkov za preverjanje praktičnega znanja pri športni vzgoji - odbojka.* [*Setting and forming the measuring criteria for evaluation of practical knowledge in physical education - volleyball*] Diplomsko delo. Ljubljana: Fakulteta za šport.

Rutar Ilc, Z. (2000). Merila ocenjevanja znanja. [*Criteria for knowledge evaluation*] *Vzgoja in izobraževanje, 31*(2-3), 77–78

Rutar Ilc, Z. (2003). *Pristopi k poučevanju, preverjanju in ocenjevanju*. [*Approaches to teaching, evaluation and assessment*] Ljubljana: Zavod Republike Slovenije za šolstvo.

Sagadin, J. (1993): *Poglavja iz metodologije pedagoškega raziskovanja*. [*Chapters from the methodology of pedagogical research*] Ljubljana: Zavod Republike Slovenije za šolstvo in šport.

Strel, J., Kovač, M., & Jurak, G. (2007). *Physical and motor development, sport activities and lifestyles of Slovenian children and youth – changes in the last few decades*. Chapter 13. In W. D. Brettschneider & R. Naul (Eds.), Obesity in Europe: Young people's physical activity and sedentary lifestyles (pp. 243–264). Sport sciences international, No. 4. Frankfurt am Main: Peter Lang.

Šimůnková, I., Novotná, V., & Chrudimsky, J. (2013). Contribution of gymnastic skills to the educational content of physical literacy in elementary school children and youth. *In Proceedings of the 9th International Conference. Sport and Quality of Life 2013* (pp. 129–137). Brno, Czech Republic: Masaryk University Campus.

Štemberger, V. (2000). Opisno ocenjevanje pri športni vzgoji v prvi triadi osnovne šole. [*Descriptive assessment of physical education in first three-year period of primary school*] V J. Krek & M. Cencič (Eds.), *Problemi ocenjevanja in devetletna osnovna šola* (pp. 277-–290). [*Evaluation problems and primary school*] Ljubljana: Pedagoška fakulteta Univerze v Ljubljani in Zavod RS za šolstvo.

Tomažin, K., Jan, I., Škof, B., Dolenc, Plavčak, M., Čoh, M., & Dragan, R.

(2001a). Model ocenjevanja atletske motorike v prvem triletju osnovne šole in njegovo preverjanje v praksi. [*A model of evaluating athletic motor skills in the first three-year stage of primary school and its monitoring in practice*] *Šport, 50*(2), 17–21.

Tomažin, K., Jan, I., Škof, B., Dolenc, Plavčak, M., Čoh, M., & Dragan, R. (2001b). *Model ocenjevanja atletske motorike v drugem triletju osnovne šole in njegovo preverjanje v praksi.* [*A model of evaluating athletic motor skills in the second three-year stage of primary school and its monitoring in practice*] *Šport, 50*(3), 12–16.

Tomažin, K., Jan, I., Škof, B., Dolenc, Plavčak, M., Čoh, M., & Dragan, R. (2001c). *Model ocenjevanja atletske motorike v tretjem triletju osnovne šole in njegovo preverjanje v praksi.* [*A model of evaluating athletic motor skills in the last three-year stage of primary school and its monitoring in practice*] Neobjavljeno delo. Ljubljana: Fakulteta za šport.

Tomažin, K., Plavčak, M., Jan, I., Škof, B., Dolenc, A., Čoh, M., Dragan, R., & Marcina, P. (2002). Primer spremljanja, vrednotenja in ocenjevanja učencev pri pouku športne vzgoje. [*An example of evaluation and assessment in physical education lessons*] V B. Škof & M. Kovač (Eds.), *Zbornik 15.strokovnega posveta športnih pedagogov Slovenije – Uvajanje novosti pri šolski športni vzgoji* (pp. 130–147). Ljubljana: Zveza društev športnih pedagogov Slovenije.

Tsui, A. (2009). Distinctive qualities of expert teachers. *Teachers and Teaching: Theory and Practice, 4*(15), 421–439.

Voglar, M. & Kovač, M. (2002). Ples na zunanjem preverjanju in ocenjevanju znanja iz športne vzgoje ob koncu devetletke. [*Dance at external evaluation and assessment of physical education at the end of nine-year primary school*] V B. Škof & M. Kovač (Eds.), *Zbornik 15. strokovnega posveta športnih pedagogov Slovenije – Uvajanje novosti pri šolski športni vzgoji* (pp. 148–153). Ljubljana: Zveza društev športnih pedagogov Slovenije.

Zadražnik, M. (2002). Zanesljivost in objektivnost ocenjevanja praktičnega znanja iz odbojke pri pouku športne vzgoje. [*Reliability and objectivity of evaluating practical volleyball skills at physical education lessons*] V R. Pišot, V. Štemberger, F. Krpač, & T. Filipčič (Eds.), *Zbornik 2. mednarodnega znanstvenega in strokovnega posveta Otrok v gibanju* (pp. 415–420). Ljubljana: Pedagoška fakulteta.

Živčić Marković, K., Sporiš, G., & Čavar, I. (2011). Initial state of motor skills in sports gymnastics among students at Faculty of Kinesiology. *Acta Kinesiologica 5*(1), 67–72.

ACKNOWLEDGEMENTS

Funding for the study has been provided by the Slovenian Research Agency (ID No. L5-6448-C). The study would not be conducted without exceptionally voluntary work of all testers (students and researchers) and all three PE teachers – external evaluators.

Corresponding author:

Matej Majerič, Ph.D.
University of Ljubljana - Faculty of sport
Gortanova 22 Ljubljana 1000
Slovenia
Tel: + 386 31 753 333
E-mail: matej.majeric@fsp.uni-lj.si