Evaluating a sports education rubric for use as an instructional guide

HIROFUMI MAEHANA¹, HIDETADA KISHI², TAKAHIRO WATANABE³, KOYA SUZUKI⁴, MAKOTO IBUKI⁵

¹Department of Psychological Counseling, Faculty of Human Sciences, Mejiro University. JAPAN.

¹Department of Health Science, Faculty of Health and Sports Science, Juntendo University. JAPAN.

^{2,3,4}Graduate School of Health and Sports Science, Juntendo University. JAPAN.

⁵Department of Childhood Education and Welfare, Faculty of Human Sciences, Mejiro University. JAPAN.

Published online: November 30, 2021

(Accepted for publication November 15, 2021)

DOI:10.7752/jpes.2021.s6425

Abstract:

Problem Statement: For the formative use of rubrics, it is important that they are designed as instructional guides and not just scoring tools. The present study explored the need for constructive alignment (of learning, instruction, and assessment) and pedagogical methods to create a generic educational guide. It also examined whether gender differences affected instructional interventions such as peer assessment and feedback. **Purpose:** To present the structure of an instructional guide that integrates constructive alignment and pedagogical methods in the application of rubrics. **Approach:** This study used a mixed-methods research design based on qualitative and quantitative research. Participants were 283 first-year university students in Sport education classes at a Japanese university (treatment class: 69 male and 75 female students and control class: 71 male and 68 female students). In this study, the rubric's design process was specified in detail, and a highly transparent scoring strategy (electronic scoring template/tool) was used in the classroom. In the treatment class, peer feedback and peer assessment were used as instructional interventions within the rubric. **Results:** (1) The presence or absence of peer feedback and peer assessment has an impact on the educational effect of the rubric. (2) The tendency of students' self-evaluations and teachers' evaluations differed according to student gender. (3) There was a relationship between high- and low-educational effects and the appropriateness of study content, appropriateness of learning instruction, and willingness to learn spontaneously, while the assessment using rubrics was found to be independent. **Conclusion:** Based on the above results, we present a model for designing rubrics that synthesizes constructive alignment and pedagogical methods in an instructional guide. The findings of this study could be used as guidelines for the design process and formative use of rubrics.

Key Words: sport education; meta-cognitive activities; educational effect; scoring strategy; electronic scoring template/tool

Introduction

The word "rubric" historically referred to the directions for church rituals included in some liturgical books. Eventually, it came to mean a code of conduct, and in the United States (US) in the 1980s, educators started using it to describe absolute valuation criteria to be used in tandem with a proscribed valuation method. In recent years, following Europe and the United States, the use of rubrics in education has been increasing in Japanese universities. Morimoto (2008) points out that changes in learning theories and evaluation have shifted over the years from behaviorism to cognitivism, constructivism, and social constructivism, as shown in Table 1.

Table 1. Learning and evaluation methods from the perspective of learning theories.

Learning theories		Behaviorism	Cognitivism	Constructivism	Social constructivism
Learning	Characteristics	Schoolized Learning		Authentic learning	
	Knowledge perspective	Knowledge is universally true		Knowledge is something that each of us constructs for ourselves	Knowledge is a construct in social work
	Outlook on learning	Knowledge transfer		Qualitative change from learner's prior knowledge to posterior knowledge	Qualitative change from learners' prior knowledge to posterior knowledge (internalization through the social work of the community)
	Main constituent	Teacher-centered		Learner-centered	

	Learners' attitudes	Passive		Active and autonomous	
	Learning task	Schoolized assignments		Authentic issues	
Evaluation	Characteristics	Schooled assessment		Authentic assessment	
	Assessment period	Specific point in time		Continuous	
	Assessment form	Objective assessment of tests		Subjective assessment of learner's performance	
	Object to be evaluated	Focus on test scores		Emphasis on learning artifacts and documentation through the process of learning activities	
	The state of assessment	Assessment disconnected from learning		Assessment embedded in learning	
	Assessment Method	Tests		Portfolios and rubrics	
		Test of someone's ability	Learning process identification and diagnostic assessment	Self-regulation Self-assessment	Peer assessment (by experts) Peer feedback (by experts)

The social constructivist approach emphasizes the importance of students applying standards through dialogue and working with both explicit and implicit knowledge of the standards (Carless & Chan, 2017; To & Liu, 2018). Performance assessment is defined by Stiggins (1987) as the evaluation of a series of attempts by learners to use the knowledge and skills they have already acquired to solve a new problem or perform a specific task. Performance assessment was originally proposed in the US in the late 1980s as a representative evaluation method in authentic evaluation theory, in response to criticism of top-down standardized testing by state governments. Performance assessment focuses on the process as well as the products and can elicit higher-level thinking and problem-solving skills (Brown & Hudson, 1998). In addition, it allows for flexibility in addressing individual needs and provides information for teaching and learning that leads to improved student performance (Valdez-Pierce, 2002).

The word rubric, as used in education, is a scoring scale or grading guide for assessing student performance along with a task-specific set of criteria. A rubric is commonly a document that articulates the expectations for an assignment and lists the criteria and other factors required, while describing the levels of quality from excellent to poor (Andrade, 2000; Reddy & Andrade, 2010). Rubrics have three features: assessment criteria, grading strategy, and standards/quality definitions (Popham, 1997), which are intended to help students reach acceptable performance standards and enable them to achieve task goals for the grades they want (Panadero & Romero, 2014). The assessor considers the rubric's criteria when determining the quality of a student's performance. The criteria may also be referred to as guidelines, as they reflect the processes and content that have been determined as important for the course of study (Parke, 2001). Quality definitions provide a detailed explanation of what a student must do to demonstrate a skill, proficiency, or criterion to attain a particular level of achievement (Reddy & Andrade, 2010). Using assessment information to inform students of their progress can reduce their anxiety about the assignment and support the formative assessment process (Black & Wiliam, 2009; Wiliam, 2011; Wolters, 2003). In addition, increasing the accuracy of self-assessment scores (Jonsson & Svingby, 2007) and self-efficacy (Andrade, Wang, Du, & Akawi, 2009; Panadero, Alonso-Tapia, & Huertas, 2012) can aid the feedback process between teachers and students (Jonsson & Panadero, 2018). In this way, rubrics are beneficial for both students and teachers, as the use of rubrics allows teachers to clarify the evaluation criteria and grade fairly (Jonsson & Svingby, 2007; Schafer, Swanson, Bené, & Newberry, 2001). Rubrics are expected to ensure transparency of evaluation, reduce learners' anxiety, support the self-assessment and feedback processes, and promote performance improvement. It is important to note that rubrics that use complex language may reduce learning activities by confusing students, and that criteria and standards/quality definitions cannot be edited after the rubric is developed, and other aspects should be kept in mind (Bacchus, Colvin, Knight, & Ritter, 2020).

Advocates for the use of rubrics for formative assessment believe that rubrics can promote student learning, as well as lead to positive changes in instruction (Jonsson & Panadero, 2018). Biggs (1996) refers to "constructive alignment" when referring to the teacher's use of a rubric to improve the alignment of learning, instruction, and assessment by clarifying the assessment criteria. The use of rubrics has been shown to improve student performance. However, the research on the use of rubrics is not unanimous. While some studies report large effects on student performance, others have found only small changes or none at all (Andrade, Du, & Mycek,

2010; Andrade, 2001; Panadero, Alonso-Tapia, & Reche, 2013). Handley and Williams (2011) mention that rubrics can be excellent learning tools, but Sadler (2005) states that they lack creative and flexible learning. Colvin et al. (2016) suggest that rubrics alone may not be effective in helping students complete tasks for assessment. Conversely, some researchers stress the instructional value of rubrics and urge instructors to use them as instructional guides, not just grading tools (Andrade, 2000; Osana & Seymour, 2004; Song, 2006; Tierney & Simon, 2004). These reports indicate that there is a need to accumulate further knowledge on how rubrics support student learning and what factors influence their success or failure.

Studies on the use of rubrics have been conducted in the liberal arts, information literacy, medicine, nursing, management, dentistry, food technology, teacher education, and film technology (Reddy & Andrade, 2010). Sport education (SE), primarily implemented in higher education, is an instructional model that links sports as taught in physical education classes to the broader sporting culture (Siedentop, 1995). Some suggest that SE provides a focus on participation in a wide range of sporting activities rather than skills acquisition. Sports education provides opportunities for pupils of all abilities to achieve in a variety of roles associated with competitive and non-competitive sports, and it highlights the importance of inclusion within the physical education environment. There are few detailed reports on the use of rubrics in the field of SE. Empirical research is needed to examine evaluation and instruction methods, such as self-regulation, self-assessment, peer feedback, peer assessment, and scoring strategies to improve constructive alignment and provide generalizable instructional guidance.

It is a major issue that most of the research on rubrics lacks sufficient assessment of the reliability and validity of the development processes for establishing rubric quality. Previous research on rubrics has been mainly qualitative, and the challenge is to promote quantitative research incorporating versatile scoring strategies. In addition, more knowledge is needed about the interaction between the use of rubrics and gender differences. In responding to the above issues, there is a need for a study using a mixed-methods research design with inter-rater reliability in the use of rubrics, questionnaires to assess intervention aspects, and qualitative and quantitative surveys of academic grades. Most of the peer-reviewed reports on rubrics have been published in the US and Australia. These follow those countries' research models, so the accumulation of international research results based on national instructional designs from other regions is still in its infancy.

The purpose of this study was to clarify the procedures in the rubric design process based on empirical research for SE in Japanese higher education, and to present an instructional guide structure that integrates constructive alignment and pedagogical methods in a rubric employing a transparent scoring strategy. The questions that motivated this study were as follows: (1) Does the combined use of rubrics and meta-cognitive activities affect learning effectiveness in SE? (2) Do gender differences influence learning that uses a combination of rubrics and meta-cognitive activities? (3) Does the increase in constructive alignment and learner motivation have a positive impact on educational effect?

Material & methods

Participants

We recruited students from SE courses from a pool of first-year students in the College of Business Administration, Psychology, Sociology, and Education at a Japanese university. Sports education courses are compulsory for these majors. The SE courses were organized into 10 classes with a capacity of 35 students, and there were 15 sessions over a period of approximately 5 months. The contents of the sessions were as follows: session 1, sport recreation; session 2, exercise/fitness; sessions 3 and 4, badminton; sessions 5 and 6, volleyball; sessions 7 and 8, football; sessions 9 and 10, basketball; sessions 11 and 12, table tennis; sessions 13 and 14, floorball; and session 15, para sports/amputee football. The 10 classes were equally divided into 5 treatment classes and 5 control classes. All the classes were conducted based on the rubrics designed for this study.

This study was conducted with the approval of the research ethics committee of the university. All participants were informed about the methods and procedures of the study. Two hundred and eighty-three students agreed to participate in the study (treatment class: 69 males and 75 females; control class: 71 males and 68 females) and gave their written consent. The participants did not include students with special needs such as visual, hearing, or physical disabilities, and no students were simultaneously taking other courses using the rubric.

Procedure

The rubric was designed in accordance with the SE subject characteristics, criteria and standards/quality definitions, and a level of achievement of 4, excellent; 3, good; 2, fair; and 1, poor. Although some reports recommend collaboration between students and teachers in designing rubrics (Reddy & Andrade, 2010), it is necessary to consider the academic course, subject characteristics, and students' grades and grade point averages. In this study, five teacher members with more than 10 years of experience teaching SE courses in higher education reviewed the evaluation process and content, and they designed a rubric with simplified content that avoided the use of complex language for describing criteria and standards/quality definitions. The rubrics were piloted four times in the SE courses. The content of the rubric was not changed after the pilot. The content of the rubric is shown in Table 2.

Table 2. Content of rubrics in Sport education

Criteria			Standards/quality definitions			
Subject	Item	Purpose	4	3	2	1
Attitude (50%)	Health management (30%)	Be able to manage their physical condition, attend classes in perfect condition, and take appropriate actions according to the living and learning environment.	One or less absences is the standard. Have an exemplary rhythm of life and exercise habits, and behave in a manner that considers health and safety management.	Less than two absences will be considered as the standard. Have a good rhythm of life and exercise habits, and act with health and safety management in mind.	Three or four absences or less is the standard. Interested in health and safety management in regular life.	Five or more absences is the standard. Non interest in health and safety management in regular life.
	Ingenuity (10%)	Understand the rules and learning methods of each discipline and be able to conduct appropriate learning activities according to the purpose.	Understand the rules and learning methods accurately and suggest ways to learn and practice.	Understand the rules and learning methods.	Interested in understanding rules and learning methods.	Non understand the rules or how to learn.
	Readiness (10%)	Clarify the purpose and goals of the session and deepen the learning process.	Be able to think about ways to deepen their learning before coming to class.	Looking for ways to deepen learning in session.	Interested in ways to deepen learning.	Non interested in ways to deepen learning
Ability (20%)	Aspiration (10%)	Be able to change and improve our behavior and abilities to achieve a better state.	Demonstrate exemplary behavior and changes in ability to co-learners.	It shows an improvement in behavior and ability to achieve a better state.	It shows that we are striving for a better state.	Non trying hard enough.
	Cooperation (10%)	Be able to think about learning methods and work with co-learners according to differences in gender, ability, and physical strength.	Demonstrate good leadership/followership and work with co-learners.	Taking the initiative in working with co-learners.	Have the attitude to work with co-learners.	Non intention of working with co-learners.
Knowledge (30%)	Explicit knowledge (30%)	Be able to logically externalize the experiences and achievements in the session and combine them with knowledge based on appropriate and precise procedures to present them as explicit knowledge.	Experience and achievements are logically externalized and combined with knowledge based on excellent procedures to show explicit knowledge.	Experiences and achievements are logically externalized and combined with knowledge to show explicit knowledge.	Experience and achievements are logically externalized.	Non able to logically externalize our experiences and achievements.

To verify the inter-rater reliability of the designed rubrics, two teachers in charge of the course were randomly selected, and a prior inspection was conducted by collecting the spreadsheets in which each item was evaluated in one of the pilot classes. The reliability of the collected data were verified using Cohen's kappa (Cohen, 1960). Cohen's coefficient values for each assessment were health management, 1; ingenuity, 0.87; readiness, 0.83; aspiration, 0.89; cooperation, 0.84; and explicit knowledge, 0.83. The results showed that the measurements were reliable (Landis & Koch, 1977).

As a scoring strategy, Campbell (2005) suggests that the use of electronic scoring tools allows for consistent, reliable, and efficient scoring. This study set up the grade calculation sheet as an electronic scoring template/tool to provide an objective basis for grade inquiries and to correct arbitrary evaluations by teachers. The format was a spreadsheet compatible with electronic tablets so that teachers could fill in student evaluations at any time in each class. For each item, the total score and academic grade (A, 100-85; B, 84-70; C, 69-50; D, ≤ 49) were automatically provided after entering the number 4, 3, 2, or 1, respectively. The designed electronic scoring template/tool is shown in Table 3.

Table 3. Electronic scoring template/tools as grade calculation sheet in Sport education.

Student information			Grade calculation sheet												Total score /100)	Academic grade
			Attitude (50%)						Ability (20%)				Knowledge (30%)			
Student ID	Name	Gender	Health manag (30%)	Score ¹	Ingenuity (10%)	Score ²	Readiness (10%)	Score ³	Aspiration (10%)	Score ⁴	Cooperation (10%)	Score ⁵	Explicit Knowledge (30%)	Score ⁶		
1000001	Example 1	Male	4	30	4	10	4	10	4	10	4	10	4	30	100	A
1000002	Example 2	Female	3	22.5	4	10	4	10	4	10	4	10	3	22.5	85	A
1000003	Example 3	Male	4	30	3	7.5	3	7.5	4	10	3	7.5	3	22.5	85	A
1000004	Example 4	Male	4	30	3	7.5	2	5	3	7.5	2	5	3	22.5	77.5	B
1000005	Example 5	Female	2	15	2	5	2	5	2	5	2	5	2	15	50	C
1000006	Example 6	Male	1	7.5	2	5	2	5	2	5	1	2.5	1	7.5	32.5	D
1000007	Example 7	Female	4	30	3	7.5	3	7.5	3	7.5	3	7.5	3	22.5	82.5	B
1000008	Example 8	Male	3	22.5	3	7.5	3	7.5	2	5	3	7.5	4	30	80	B
1000009	Example 9	Female	3	22.5	3	7.5	3	7.5	3	7.5	3	7.5	4	30	82.5	B
:	:	:		0		0		0		0		0		0	0	D
:	:	:		0		0		0		0		0		0	0	D

Score (programming code) : Score¹ (=D4/4*100*0.3) ,Score² (=F4/4*100*0.1) ,Score³ (=H4/4*100*0.1) ,Score⁴ (=J4/4*100*0.1) ,Score⁵ (=L4/4*100*0.1) ,Score⁶ (=N4/4*100*0.2) ,Total score (=SUM(E4,G4,I4,K4,M4,O4)) .
Academic grade (programming code) : A; 100-85, B; 84-70, C; 69-50, D; ≤ 49 (=IF(\$P4>=85,"A",IF(\$P4>=70,"B",IF(\$P4>=50,"C","D")))) .

Data collection and analysis

In the initial guidance, the significance and characteristics of rubrics in SE were explained, and guidance was given on self-regulation and self-assessment efforts using the rubric. After the eighth session, the teachers calculated the tentative grade information as the students' midpoint scores using the electronic scoring template/tool. The critical differences between the treatment class and the control class are as follows. In the treatment class only, peer feedback using the rubric was given to the students by the teachers at the end of the eighth session, and information on their performance during the course was disclosed, as well as advice on how to improve future efforts as indicated by the peer assessments. At the end of the fifteenth session, students in both the control and treatment classes calculated their expected grades by using the electronic scoring template/tool for self-assessment and submitted the results (self-assessment score) to the teacher. In the same way, the teachers also disclosed their grades (total score) to the students using the electronic scoring template/tool.

Based on Florence's (2006) study of the educational effects of a rubric-based class, we created a questionnaire with responses on a Likert-type scale (1, not at all; 2, a little; 3, somewhat; 4, very; 5, extremely). After all the work in session 15 was completed, students were asked to respond to the questionnaire. Encryption was used to maintain the anonymity of the respondents and to enable matching of the respondents with the response forms.

The questions were as follows. Question 1/Learning: Was the content of learning (quality and quantity of learning) appropriately in accordance with the rubrics? Question 2/Instruction: Was the teacher's learning guidance appropriately in accordance with the rubrics? Question 3/Assessment: Were the evaluations of the rubric criteria, standards/quality definitions, and scoring strategy (electronic scoring template/tool) appropriate? Question 4/Learner's motivation: Has the rubric increased your active and autonomous motivation for learning? The above research protocol is shown in Figure 1.

Figure 1. Overview of research protocols in treatment class and control class. Dimgrey inverted triangle for teacher's outcome and white triangle for student's outcome.

The Welch-Aspin test was used to compare the total scores, midpoint scores, self-assessment scores, and questionnaire answers in the treatment and control classes. In the comparison of four groups, Tukey's honestly significant difference test was used for the multiple comparison together with a one-way analysis of variance. The Wilcoxon signed-rank test was used for within-group comparisons, and Mann-Whitney's U test was used to compare total and self-assessment scores. Pearson's correlation analysis was used to evaluate the relationship between total and self-assessment scores, and the total score and questionnaire answers. All variables are presented as mean $\pm$ standard deviations. SPSS version 27.0.1 (IBM, Chicago) was used for the calculations, and statistical significance was inferred at $p < 0.05$.

Results

The results of the comparison of total, midpoint, and self-assessment scores in the treatment and control classes are as follows. The total score of the treatment class 83.70 ± 6.51 was significantly higher than that of the control class at 77.41 ± 6.73 ($p < 0.01$). The midpoint score for the treatment class was 76.84 ± 6.79 and for the control class, 76.91 ± 7.32 . The self-assessment score for the treatment class was 80.89 ± 7.51 and for the control class, 79.41 ± 7.43 , and there was no significant difference between the classes in both items.

The results of the comparison of midpoint and total scores for each of the treatment and control classes are as follows. In the treatment class, there was a significant difference between the midpoint score, 76.84 ± 6.79 , and total score, 83.70 ± 6.51 ($p < 0.01$). In the control class, there was no significant difference between the midpoint score, 76.91 ± 7.32 , and total score, 77.41 ± 6.73 .

The results of the comparison of total and self-assessment scores for each of the treatment and control classes are as follows. In the treatment class, the total score 83.70 ± 6.51 and was higher than the self-assessment score 80.89 ± 7.51 , and there was a significant difference ($p < 0.01$). In the control class, the self-assessment score 79.41 ± 7.43 was higher than the total score 77.41 ± 6.73 , and there was a significant difference ($p < 0.01$).

The correlation coefficients between total and self-assessment scores for each class were treatment class ($r = 0.53$, $p < 0.01$) and control class ($r = 0.36$, $p < 0.05$), respectively.

Figure 2 shows a comparison of total scores for treatment class male students (TC male), treatment class female students (TC female), control class male students (CC male), and control class female students (CC female). TC female 85.57 ± 5.92 was the highest value among the four groups, and CC male 74.82 ± 6.54 was the lowest value among the four groups. TC female was significantly higher than TC male, CC male, and CC female ($p < 0.01$). TC male 81.67 ± 6.55 was significantly higher than CC male ($p < 0.01$). CC female at 80.11 ± 5.85 was also significantly higher than CC male ($p < 0.01$).

Figure 2. Influence of gender difference on total score by multiple comparisons of four groups. Dimgrey bar line; treatment class male student (TC male), black bar line; treatment class female student (TC female), lightgrey bar line; control class male student (CC male) and silver bar line; control class female student (CC female). Significant difference (**; $p < 0.01$).

The midpoint and total scores for TC male, TC female, CC male, and CC female are shown in Figure 3. In the treatment class, there was a significant difference between the midpoint scores (TC male, 76.96 ± 7.21 ; TC female, 76.73 ± 6.43) and the total scores (TC male, 81.67 ± 6.55 ; TC female, 85.57 ± 5.92) for both male and female students ($p < 0.01$). In the control class, there was no significant difference between the midpoint scores (CC male, 74.73 ± 8.19 ; CC female, 79.19 ± 5.47) and the total scores (CC male, 74.82 ± 6.54 ; CC female, 80.11 ± 5.85) for both male and female students.

Figure 3. Comparison of midpoint score and total score in four groups. Black dotted line; treatment class male student (TC male), black line; treatment class female student (TC female), grey dotted line; control class male student (CC male) and grey line; control class female student (CC female). Significant difference (**; $p < 0.01$).

The results of the comparison of total and self-assessment scores for TC male, TC female, CC male, and CC female are shown in Figure 4. There was no significant difference between total score at 81.67 ± 6.55 and self-assessment score at 82.43 ± 7.66 for TC male. TC female showed a significant difference between total score at 85.57 ± 5.92 and self-assessment score at 79.48 ± 7.13 ($p < 0.01$). CC male showed a significant difference between total score at 74.82 ± 6.54 and self-assessment score at 82.57 ± 7.42 ($p < 0.01$). CC female showed a significant difference between total score at 80.11 ± 5.85 and self-assessment score at 76.10 ± 5.88 ($p < 0.01$).

Figure 4. Comparison of total score and self-assessment score in four groups. Black bar line; total score and white bar line; Self-assessment score. TC male; treatment class male student, TC female; treatment class female student, CC male; control class male student and CC female; control class female student. Significant difference (; $p < 0.01$).**

The correlation coefficients between total and self-assessment scores for both classes by gender were TC male ($r=0.61$, $p < 0.01$), TC female ($r=0.66$, $p < 0.01$), CC male ($r=0.63$, $p < 0.01$), and CC female ($r=0.66$, $p < 0.01$), respectively.

The comparison of the questionnaire answers in the treatment and control classes are shown in Table 4. The treatment class showed significantly higher values in learning (4.06 ± 0.47 vs 3.91 ± 0.48), instruction (4.40 ± 0.68 vs 3.55 ± 0.59), and learner's motivation (4.27 ± 0.68 vs 3.69 ± 0.52) than the control class ($p < 0.01$). In the assessment, there was no significant difference between the treatment class 4.10 ± 0.53 and the control class 4.09 ± 0.72 .

The correlations between total scores and the questionnaire answers in the treatment and control classes are shown in Table 5. The correlations between the total scores and the questionnaire answers in the treatment class are learning ($r=0.70$, $p < 0.01$), instruction ($r=0.77$, $p < 0.01$), assessment ($r=0.02$), and learner's motivation ($r=0.72$, $p < 0.01$), respectively. The correlation between the total scores and the questionnaire answers in the control class are learning ($r=0.03$), instruction ($r=0.45$, $p < 0.05$), assessment ($r=0.12$), and learner's motivation ($r=0.16$), respectively.

Table 5. Correlation between total score and question form in treatment class and control class.

Variables	Constructive alignment			Learner's motivation
	Learning	Instruction	Assessment	
Treatment class				
Total score	0.70**	0.77**	0.02	0.72**
Control class				
Total score	0.03	0.45*	0.12	0.16

Significant difference (* : $p < 0.05$, ** : $p < 0.01$).

Discussion

The present study used empirical research to construct a SE instructional guide integrating constructive alignment and pedagogical methods in the application of rubrics incorporating a highly transparent scoring strategy (electronic scoring templates/tools) and to specify the rubric's design process in detail. In addition, it is the only empirical study to examine the influence of gender differences in learning with a rubric. The main findings of this study are as follows. (1) The presence or absence of peer feedback and peer assessment has an impact on educational effect. (2) The tendencies of students' self-evaluation and teachers' evaluations differed by student gender. (3) There was a relationship between high- and low-educational effects and appropriateness of study content, appropriateness of learning instruction, and willingness to learn spontaneously, while the assessment using the rubric was found to be independent. The above findings were integrated into a model for instructional guides.

Many studies have shown that physical activity is an effective way to prevent age-related diseases and improve quality of life (Bize, Johnson, & Plotnikoff, 2007). Sports education is an important subject that can inspire students' interest in their health and exercise habits. However, it has been reported that Japanese university students exhibit significantly lower levels of physical activity compared to other age groups and are less physically active than students in other countries (Caspersen, Pereira, & Curran, 2000; Haase, Steptoe, Sallis, & Wardle, 2004). It has also been reported that female students are less likely to play sports and have less familiarity with sports than male students (Maehana, Kishi, & Ibuki, 2021). In consideration of this background, the SE classes in this study actively tried to promote changes in the students' consciousness of exercise and sport that could be implemented to create exercise habits in both male and female students.

This study used a narrow perspective to position a student's total score as an "educational effect." The study findings demonstrate the educational effect of combining a rubric with meta-cognitive activities. The results of comparing the total scores in the treatment class and the control class were that the educational effect of the treatment class was superior to that of the control class. The results of the comparison of midpoint and total scores for each class suggest that the presence or absence of peer feedback and peer assessment (treatment class: presence; control class: absent) had affected the educational effect. Previous studies have not drawn clear conclusions about which factors improve student performance in formative use of rubrics (Jonsson & Svingby 2007). However, many of the studies that examined these factors employed a research design that combined the use of rubrics with other instructional interventions. Wiliam (2011) mentions that using rubric-based assessment information to inform students of their progress and to support the formative assessment process helps students grow and influences their performance.

This study showed that peer feedback and peer assessment employed in the treatment classes increased learning effectiveness. However, there is no evidence that the use of rubrics in and of themselves can lead to improved educational effects. It is also possible that improvements in student performances were merely conversions to the performances desired by their teachers. In addition, it is necessary to consider the possibility that students' free trial and error processes may be hindered in part, and their creativity may be reduced under the influence of rubrics. For convenience, the self-assessment scores were replaced by the students' self-evaluations and teachers' total evaluations in the discussion of assessment. As a result of comparing the students' self-evaluations and the teachers' evaluations for each class, it became clear that the trends in the two classes were different. In international comparative studies on self-esteem and self-affirmation, it is often pointed out that Japanese youth have low levels of self-affirming emotions. Schmitt and Allik (2005) pointed out that Japanese people had the lowest self-esteem in a study of 53 countries. Conversely, research on latent self-esteem has shown that Japanese people do not have low self-esteem (Yamaguchi et al., 2007). The low self-esteem and self-affirming emotions of Japanese adolescents found in international comparisons are most likely due to the way that people present themselves in Japanese culture or the measurement methods that were used. In Japanese society, it is necessary to consider the cultural context in which humility is seen as a virtue. In other words, the students' self-evaluations may have reflected the scores that they felt demonstrated humility, making them more likely to be accepted and admired by others. Therefore, to improve the consistency of teacher and student evaluations while pursuing the educational effects achieved by using rubrics and meta-cognitive activities together, it is necessary to include evaluation models of self-esteem and self-affirmation.

The results of the comparison of the total scores of male and female students in both classes showed that the total scores of female students were superior to that of male students. The results of the comparison of midpoint and total scores according to gender also showed that female students tended to gain a better educational effect from peer feedback and peer assessment than male students. In both classes, the male students' self-evaluations were higher than their teachers' evaluations. On the contrary, the teachers' evaluations were higher than their female students' self-evaluations. Therefore, we suggest that the tendency of student self-evaluations and teacher evaluations differs according to student gender. Fostering self-regulation and self-efficacy using rubrics is a pedagogical technique to promote students' monitoring and self-evaluation strategies, but few studies have examined rubrics considering gender differences. Efklides (2011) refers to self-regulation as the ability to appropriately assess one's own knowledge and skills and plan one's actions according to learning content and tasks, and reports that self-regulation has a high affinity with the use of rubrics.

From the comparison of the total scores of male and female students in both classes, female students were better than male students at self-regulation using the rubric. Therefore, as Panadero and Jonsson (2013) mention, there may be a gender difference influence on students' clearer understanding of what they need to achieve in learning and how they are expected to perform. Self-efficacy is an individual's belief in the ability to perform certain actions necessary to achieve the desired result and has been reported to be associated with physical activity (Luszczynska & Haynes, 2009; Blake, Stanulewicz, & McGill, 2017; Hamilton, Warner, & Schwarzer, 2017). Pajares (2008) reported that students with high self-efficacy perform better than those with low self-efficacy. Dinther et al. (2010) point out that students need to receive feedback from their teachers in order to gain an objective understanding of their own performance. In this study, peer feedback and peer assessment were

employed in the treatment classes. The comparison of midpoint and total scores showed that female students tended to experience a higher educational effect than male students. Andrade et al. (2009) state that the use of rubrics in combination with feedback can enhance self-efficacy. This study did not measure self-efficacy prior to the instructional intervention and did not show evidence that female students have higher self-efficacy development than male students. However, we are able to present findings that suggest examination of the methodology of instructional interventions based on the influence of gender differences in learning effectiveness and self-efficacy. On the other hand, for male students in the treatment class, there were no significant differences between students' self-evaluations and teachers' evaluations. Previous studies have reported that lower achievers tend to overestimate themselves and higher achievers tend to underestimate themselves (Landrum, 1999; Lejk & Wyvill, 2001). Our novel findings suggest that the use of rubrics and meta-cognitive activities together mitigated the difference in scores between students' self-evaluations and teachers' evaluations.

This study examined constructive alignment (of learning, instruction, and assessment) and motivation to learn as measured by educational effect. Comparison of the questionnaire answers from the test and control classes found that there was a significant difference in learning $\equiv$ appropriateness of study content, instruction $\equiv$ appropriateness of learning instruction, and learner's motivation $\equiv$ willingness to learn spontaneously, and that the treatment class was superior to the control class. However, it is also important to note that the assessment did not show a significant difference between classes.

We examined the effects of instructional interventions, an issue also raised in previous research, from the perspectives of constructive alignment and motivation to learn. Regarding constructive alignment, we found an effect on appropriateness of study content and appropriateness of learning instruction. In addition, willingness to learn spontaneously was also found to have a positive effect. However, our assessment of constructive alignment did not show any effect from the instructional interventions. The relationship between constructive alignment and a learner's motivation as measured by high and low total scores in both classes showed that there was no relationship between the scores and assessment in both classes. This is a useful finding as it shows that rubrics are independent of educational effects.

Factors that helped to ensure the independence of the assessment included employing a transparent scoring strategy, an electronic scoring template/tool. Campbell (2005) notes that electronic scoring tools enable consistent, reliable, and efficient scoring. The use of a transparent scoring strategy should be standard when applying rubrics. Previous research on the usefulness of rubrics has tended to focus on the lack of evaluation differences between teachers and students. However, as shown in the results of this study, evaluation differences between teachers and students are not necessarily negative aspects. Even if a rubric is constructed that is easy for students and teacher to understand and handle, it cannot be said to be truly useful unless independence of assessment is established. Paradoxically, if we can build a rubric that establishes the independence of assessment, we can present objective evidence when questions about grades arise. Thus, a rubric that is independent of assessment is strongly recommended for higher education.

The degree of agreement between student and teacher evaluation results has been previously reported using correlation coefficients (Campbell, Mothersbaugh, Brammer, & Taylor, 2001; Lin, Liu, & Yuan, 2001; Rudy, Fejfar, Griffith, & Wilson, 2001; Patri, 2002). In line with previous studies, this study also showed the correlation between total and self-assessment scores: Significant correlations were found between classes and gender differences. Ryan et al. (2007) pointed out that a high correlation coefficient does not necessarily mean that the results of student and teacher assessments are close. Therefore, it is hoped that this research will only be reported as a reference value and will be used in future research focused on evaluation.

This study aimed to present an instructional guide structure that integrates constructive alignment and pedagogical methods for the application of rubrics. The following findings also demonstrate the quintessence of this study. Previous studies emphasized the educational value of rubrics and recommended their use as instructional guides rather than just scoring tools (Andrade, 2000; Osana & Seymour, 2004; Tierney & Simon, 2004; Song, 2006).

An instructional guide structure that integrates constructive alignment and pedagogical methods in a rubric based on the findings presented in this study is shown in Figure 5. Constructive alignment is the origin, the Y-axis is learning, the X-axis is instruction, the Z-axis is the learner's motivation, and assessment is the base plane. In learning, it is important to use meta-cognitive activities such as self-regulation and self-assessment. In instruction, it is important to use meta-cognitive activities such as peer feedback and assessment. For a learner's motivation, it is important to foster their activities and autonomy. Finally, to ensure the independence of the assessment, it is important to establish the rubric's own criteria, standards/quality definitions, scoring strategy, reliability, and validity. A harmonious expansion of these factors will increase educational effect, which in turn has a ripple effect on the development of self-efficacy. It is hoped that future research will accumulate knowledge based on the this instructional guide model.

Figure 5. Configuration diagram of an instructional guide that integrates constructive alignment and pedagogical methods in rubric operation.

The limitations of this study should be noted. First, individual differences such as goal orientation, past academic performance, and self-efficacy of students who participated in the study have not been clarified. Second, we could not establish a control group that did not use the rubric, as the ethics review board disallowed it. Third, specific peer feedback and peer assessment could not be centralized due to the diversity of students' educational needs. Therefore, a comprehensive research design needs to be established for future studies. The originality of this study lies in its use of a mixed-methods research design based on qualitative and quantitative surveys to formulate a model for the instructional guide.

Conclusions

This study showed the following findings. (1) The presence or absence of peer feedback and peer assessment has an impact on the educational effect of a rubric. (2) The tendency of students' self-evaluations and teachers' evaluations differed by student gender. (3) There was a relationship between high and low educational effects and appropriateness of study content, appropriateness of learning instruction, and willingness to learn spontaneously, while the assessment using rubrics was found to be independent. Based on the above results, we present a model for an instructional guide for a rubric based on constructive alignment and pedagogical methods.

Conflicts of interest

The authors declare no conflicts of interest.

References:

- Andrade, H. (2000). Using rubrics to promote thinking and learning. *Educational Leadership*, 57(5), 13–18.
- Andrade, H., Du, Y., & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students' writing. *Assessment in Education: Principles, Policy and Practice*, 17(2), 199–214.
- Andrade, H.G. (2001). The effects of instructional rubrics on learning to write. *Current Issues in Education*, 4(4).
- Andrade, H., Wang, X. L., Du, Y., & Akawi, R. L. (2009). Rubric-referenced self-assessment and self-efficacy for writing. *Journal of Educational Research*, 102(4), 287–301.
- Bacchus, R., Colvin, E., Knight, E. B., & Ritter, L. (2020). When rubrics aren't enough: Exploring exemplars and student rubric co-construction. *Journal of Curriculum and Pedagogy*, 17 (1), 48–61.
- Biggs, J. (1999). *Teaching for quality learning at University*. Buckingham: SRHE and Open University Press, 165–203.
- Bize, R., Johnson, J. A., & Plotnikoff, R. C. (2007). Physical activity level and health-related quality of life in the general adult population: a systematic review. *Preventive Medicine*, 45, 401–415.
- Black, P., & William, D. (2009). Developing the theory of formative assessment. *Educational Assessment Evaluation and Accountability*, 21, 5–31.
- Blake, H., Stanulewicz, N., & McGill, F. (2017). Predictors of physical activity and barriers to exercise in nursing and medical students. *Journal of Advanced Nursing*, 73, 917–929.
- Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, 32 (4), 653–675.
- Campbell, A. (2005). Application of ICT and rubrics to the assessment process where professional judgment is involved: the features of an e-marking tool. *Assessment and Evaluation in Higher Education*, 30(5), 529–37.

- Campbell, K. S., Mothersbaugh, D. L., Brammer, C., & Taylor, T. (2001). Peer versus self-assessment of oral business presentation performance. *Business Communication Quarterly*, 64(3), 23–40.
- Carless, D., & Chan, K. K. H. (2017). Managing dialogic use of exemplars. *Assessment and Evaluation in Higher Education*, 42(6), 930–941.
- Caspersen, C. J., Pereira, M. A., & Curran, K. M. (2000). Changes in physical activity patterns in the United States, by sex and cross-sectional age. *Medicine and Science in Sports and Exercise*, 32(9), 1601–1609.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.
- Colvin, E., Bacchus, R., Knight, E., & Ritter, A. (2016). Exploring the way students use rubrics in the context of criterion referenced assessment. 39th annual conference of the Higher Education research and Development Society of Australasia, 39(11), 42–52.
- Dinther, M. V., Dochy, F., & Segers, M. (2010). Factors affecting students' self-efficacy in higher education. *Educational Research Review*, 6(2), 95–108.
- Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational Psychologist*, 46(1), 6–25.
- Florence, J. (2006). Rubrics for teacher education in community college. *the Community College Enterprise*, 39–55.
- Haase, A., Steptoe, A., Sallis, J. F., & Wardle, J. (2004). Leisure-time physical activity in university students from 23 countries: Associations with health beliefs, risk awareness, and national economic development. *Preventive Medicine*, 39, 182–190.
- Hamilton, K., Warner, L. M., & Schwarzer, R. (2017). The Role of Self-Efficacy and Friend Support on Adolescent Vigorous Physical Activity. *Health Education and Behavior*, 44, 175–181.
- Handley, K., & Williams, L. (2011). From copying to learning: Using exemplars to engage students with assessment criteria and feedback. *Assessment and Evaluation in Higher Education*, 36(1), 95–108.
- Jonsson, A., & Panadero, E. (2018). Facilitating Students' Active Engagement with Feedback. In A. Lipnevich & J. Smith (Eds.), *The Cambridge Handbook of Instructional Feedback* (Cambridge Handbooks in Psychology), 531–553.
- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2, 130–144.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement of categorical data. *Biometric*, 33, 159–174.
- Landrum, R. E. (1999). Student expectations of grade inflation. *Journal of Research and Development in Education*, 32, 124–128.
- Lejk, M., & Wyvill, M. (2001). The effect of the inclusion of self-assessment with peer assessment of contributions to a group project: A quantitative study of secret and agreed assessments. *Assessment and Evaluation in Higher Education*, 26, 551–561.
- Lin, S. S. J., Liu, E. Z. F., & Yuan, S. M. (2001). Web- based peer assessment: Feedback for students with various thinking-styles. *Journal of Computer Assisted Learning*, 17(4), 420–432.
- Luszczynska, A., & Haynes, C. (2009). Changing nutrition, physical activity and body weight among student nurses and midwives: Effects of a planning intervention and self-efficacy beliefs. *Journal of Health Psychology*, 14, 1075–1084.
- Maehana, H., Kishi, H., & Ibuki, M. (2021). Gender Differences, Learning Status and Lifestyle Habits in On-Demand Distribution of Exercise Tasks. *Mejiro University Reserch Institute for Higher Education*, 27, 85–91.
- Morimoto, Y. (2008). E-Portfolios: Theory and Practice. *The Journal of Information and Systems in Education*, 25 (2), 245–263.
- Osana, H.P., & J.R. Seymour. (2004). Critical thinking in preservice teachers: A rubric for evaluating argumentation and statistical reasoning. *Educational Research and Evaluation*, 10, 473–498.
- Pajares, F. (2008). Motivational role of self-efficacy beliefs in self-regulated learning. In D. H. Schunk & B. J. Zimmerman (Eds.), *Motivation and self-regulated learning. Theory, research and applications*. New York: Lawrence Erlbaum Associates, 111–168.
- Panadero, E., Alonso-Tapia, J., & Huertas, J. A. (2012). Rubrics and self-assessment scripts effects on self-regulation, learning and self-efficacy in secondary education. *Learning and Individual Differences*, 22(6), 806–813.
- Panadero, E., Alonso-Tapia, J., & Reche, E. (2013). Rubrics vs. self-assessment scripts effect on self-regulation, performance and self-efficacy in pre-service teachers. *Studies in Educational Evaluation*, 39, 125–132.
- Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, 9, 129–144.
- Panadero, E., & Romero, M. (2014). To rubric or not to rubric? The effects of self-assessment on self-regulation, performance and self-efficacy. *Assessment in Education: Principles, Policy and Practice*, 21(2), 133–148.

- Parke, C. S. (2001). An approach that examines sources of misfit to improve performance assessment items and rubrics. *Educational Assessment*, 7(3), 201–225.
- Patri, M. (2002). The influence of peer feedback on self- and peer-assessment of oral skills. *Language Testing*, 19(2), 109–131.
- Popham, W. J. (1997). What's wrong – and what's right – with rubrics. *Educational Leadership*, 55(2), 72–75.
- Reddy, Y., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, 35, 435–448.
- Rudy, D. W., Fejfar, M. C., Griffith, C. H., & Wilson, J. F. (2001). Self- and peer assessment in a first-year communication and interviewing course. *Evaluation & the Health Professions*, 24(4), 436–445.
- Ryan, G. J., Marshall, L. L., Porter, K., & Jia, H. (2007). Peer, professor and self-evaluation of class participation. *Active Learning in Higher Education*, 8(1), 49–61.
- Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment and Evaluation in Higher Education*, 30(2), 175–194.
- Schafer, W. D., Swanson, G., Bené, N., & Newberry, G. (2001). Effects of teacher knowledge of rubrics on student achievement in four content areas. *Applied Measurement in Education*, 14(2), 151–170.
- Schmitt, D.P. & Allik, J. (2005). Simultaneous administration of the Rosenberg Self-Esteem Scale in 53 Nations: Exploring the universal and culture specific features of global self-esteem. *Journal of Personality and Social Psychology*, 89, 623–642.
- Siedentop, D. (1995). Improving sport education, Australian Council for Health, Physical Education, and Recreation Healthy Lifestyles Journal, 42, 22–24.
- Song, K.H. (2006). A conceptual model of assessing teaching performance and intellectual development of teacher candidates: A pilot study in the US. *Teaching in Higher Education*, 11(2), 175–190.
- Stiggins, R. J. (1987). Design and Development of Performance Assessments. *Educational Measurement: Issues and Practice*, 6(3), 33–42.
- Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research & Evaluation*, 9, 1–7.
- To, J., & Liu, Y. (2018). Using peer and teacher-student exemplar dialogues to unpack assessment standards: Challenges and possibilities. *Assessment and Evaluation in Higher Education*, 43(3), 449–460.
- Valdez-Pierce, L. (2002). Performance-based assessment: Promoting achievement for English language learners. *ERIC Clearinghouse on Language and Linguistics*, 26(1), 1–5.
- Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, 37, 2–14.
- Wolters, C. A. (2003). Regulation of motivation: Evaluating an underemphasized aspect of self-regulated learning. *Educational Psychologist*, 38(4), 189–205.
- Yamaguchi, S., Greenwald, A. G., Banaji, M. R., Murakami, F., Chen, D., Shiomura, K., Kobayashi, C., Cai, H., & Krendl, A. (2007). Apparent universality of positive implicit self-esteem. *Psychological Science*, 18, 498–500.