

A Comparative Study of RTOP and Chinese Classical Classroom Teaching Evaluation Tools

Dazhen Tong^{1*}, Hongjun Xing²

¹ East China Normal University, College of Teacher Education, Shanghai 200062, China

² Capital Normal University, College of Teacher Education, Beijing 100048, China

* Corresponding Author: tongdazhen1993@sina.com

(Received: 08/31/2020; Accepted: 04/06/2021; Published: 06/16/2021)

DOI: <https://doi.org/10.37906/real.2021.2>

Abstract: The new round of basic education curriculum reform in China recently will inevitably bring about the reform of classroom teaching. To find a method of testing the reformed classroom teaching effect so that it follows the new curriculum reform is a problem that needs to be solved urgently. On the basis of examining the evaluation concept and application of the "Reformed Classroom Teaching Observation Protocol (RTOP)" originated in the United States, we test the appropriateness of RTOP in China's classroom teaching evaluation environment from both qualitative and quantitative perspectives, with good results. By comparing the advantages and disadvantages of RTOP and China's "Toshiba Cup Scoring Table", it is found that RTOP can provide guidance for the transformation of classroom teaching evaluation concepts and the development of evaluation tools in the context of China's new curriculum reform.

Keywords: curriculum reform; classroom teaching evaluation; RTOP; Toshiba Cup score table

1. Introduction

In recent years, a new round of elementary education curriculum reform has been launched in China. One of the key characteristics of the new curriculum reform is the change in curriculum goals, that is, from "three-dimensional goals" (knowledge and skills, processes and methods, attitudes and values) to "core literacy". The change in curriculum goals will inevitably bring about changes in students' learning styles and teachers' teaching methods, which will lead to changes in the entire classroom teaching (Kulm & Malcom, 1991). Then the question becomes: how to evaluate the classroom teaching after the new curriculum reform to test its effect?

It is found that the United States also faced similar problems at the end of the last century. In order to make classroom teaching meet the requirements of the forthcoming (1996) *National Science Education Standards*, the National Science Foundation funded a five-year major project at Arizona State University in 1995, called "Arizona Collaborative for Excellence in the Preparation of Teachers (ACEPT)". ACEPT decided to change the way of teaching science and mathematics teachers to solve the problems in American classroom teaching, and advocated that they should be taught in the way of constructivism and inquiry. These science and mathematics teachers who have received constructivism and inquiry education will educate their students in the same way after graduation, thus bringing science and mathematics education into a virtuous cycle. In order to evaluate the effectiveness of the new teaching model that is implemented in classroom teaching, the Evaluation Facilitation Group (EFG) of ACEPT developed a classroom observation tool called "Reformed Teaching Observation Protocol (RTOP)". After several rounds of tests, it

is found that the reliability (Adamson, et al., 2003) and prediction validity (Sawada, Piburn, Judson, Turley, & Bloom, 2010) of RTOP have reached a very high level. Once RTOP was released, it was favored by educators and widely used (MacIsaac, & Dan, 2002; Amrein-Beardsley, & Popp, 2012; Campbell, & Todd, 2012; Addy, & Blanchard, 2010; Erdogan, Campbell, & Abd-Hamid, 2011). In addition, when evaluating the effectiveness of ACEPT, it is found that RTOP score has a very significant positive correlation with students' concept gain. Therefore, it shows that the teaching after reform is very effective.

For the study of classroom teaching evaluation, Chinese education researchers have long existed two tendencies, one is empiricism; the other is theoreticalism (Liu, & Xu, 2018; Xing, & Tian, 2020). The classroom teaching evaluation tools under the first tendency have such phenomena as lack of theoretical guidance, fuzzy objectives, and overlapping evaluation indicators. The classroom teaching evaluation tools under the second tendency can only provide value orientation in concept, but cannot provide an operational framework for classroom teaching evaluation. Therefore, there are few classroom teaching evaluation tools that are based on theory and have been tested in practice. In China, Toshiba (China) Co., Ltd. and the Ministry of Education have jointly held the "Toshiba Cup • Teaching Skills Innovation Competition for Science Normal University students" in order to cultivate outstanding and innovative science teachers (mathematics, physics and chemistry) since 2008. The participants are asked to design their own original teaching plans so that the primary and middle school students understand that science is all around us and enjoy the fun brought by science in teaching. The competition has attracted more than 80,000 students from 45 normal colleges and universities across the country, making it the most important teaching skills competition for pre-service teachers in China (Pan, 2012). Therefore, the "Toshiba Cup • Teaching Skills Innovation Competition for Science Students of China Normal University" score table (hereinafter referred to as "Toshiba Cup score table") can well represent China's classic classroom teaching evaluation tools.

It can be seen that RTOP and the "Toshiba Cup score table" have many similarities in terms of development purposes, applicable disciplines and populations, which makes it fair to compare the two. In addition, due to the lack of empirical research on classroom teaching evaluation in the context of China's new curriculum reform, classroom teaching urgently needs evaluation tools that can guide classroom reform. It is hoped that through the comparative study of the two classroom teaching evaluation tools, the Toshiba Cup score sheet or the RTOP evaluation sheet can be modified to promote the reform of classroom teaching in China in context of the new curriculum reform. Based on the principle of "an advice of others can serve to correct our faults", we try to translate RTOP into Chinese and use it to evaluate classroom teaching in China in order to solve the following questions:

- RTOP has good reliability and validity in American tests and research. How appropriate is it in the Chinese classroom teaching evaluation environment?
- What are the similarities and differences between RTOP and Chinese classic classroom teaching evaluation tools, and what can it provide to the improvement or development of classroom teaching evaluation tools under the concept of China's new curriculum reform?

As for the first question, we have proved that RTOP has a better suitability in the context of classroom teaching evaluation in China (Tong, Xing, & Zheng, 2020). In this article, we mainly focus on the second issue.

2. Methods and Design

2.1. Methods

In order to solve the above research problems, we adopt the following two research methods:

The first method is the case study method. A video of teacher Z's open class is selected as the research subject. The evaluators and the instructor use RTOP and Toshiba Cup score table to evaluate this class, and comprehensively discuss the improvement direction and measures of this class. This provides a good case for using RTOP to improve classroom teaching in China.

The second method is the interview method. In order to gain insights into the evaluators' views on the advantages and disadvantages of RTOP and Toshiba Cup score table, and their suggestions on the improvement or development of Chinese classroom teaching evaluation tools, we randomly select some evaluators to conduct interviews. In addition, in order to further understand the teacher's understanding of the RTOP and Toshiba Cup score table, and deeply explore the teaching reflection content of the teacher after self-evaluation, we conduct in-depth interviews with the instructor based on the evaluator's evaluation suggestions.

2.2. Tools

Two kinds of classroom teaching evaluation tools are needed in this study: RTOP scale (as shown in Table 1) and Toshiba Cup score table (as shown in Table 2).

Table 1.RTOP scale

Evaluation content	Evaluation index	Value					Score
		Inconsistent Consistent					
LESSON DESIGN AND IMPLEMENTATION (20 points)	1. The instructional strategies and activities respected students' prior knowledge and the preconceptions inherent therein.	0	1	2	3	4	
	2. The lesson was designed to engage students as members of a learning community.	0	1	2	3	4	
	3. In this lesson, student exploration preceded formal presentation.	0	1	2	3	4	
	4. This lesson encouraged students to seek and value alternative modes of investigation or of problem solving.	0	1	2	3	4	
	5. The focus and direction of the lesson was often determined by ideas originating with students.	0	1	2	3	4	
Propositional knowledge (20 points)	6. The lesson involved fundamental concepts of the subject.	0	1	2	3	4	
	7. The lesson promoted strongly coherent conceptual understanding.	0	1	2	3	4	

CONTENT (40 points)		8. The teacher had a solid grasp of the subject matter content inherent in the lesson.	0	1	2	3	4	
		9. Elements of abstraction (i.e., symbolic representations, theory building) were encouraged when it was important to do so.	0	1	2	3	4	
		10. Connections with other content disciplines and/or real world phenomena were explored and valued.	0	1	2	3	4	
	Procedural Knowledge (20 points)	11. Students used a variety of means (models, drawings, graphs, concrete materials, manipulatives, etc.) to represent phenomena.	0	1	2	3	4	
		12. Students made predictions, estimations and/or hypotheses and devised means for testing them.	0	1	2	3	4	
		13. Students were actively engaged in thought-provoking activity that often involved the critical assessment of procedures.	0	1	2	3	4	
		14. Students were reflective about their learning.	0	1	2	3	4	
	15. Intellectual rigor, constructive criticism, and the challenging of ideas were valued.	0	1	2	3	4		
CLASSROOM CULTURE (40 points)	Communicative Interactions (20 points)	16. Students were involved in the communication of their ideas to others using a variety of means and media.	0	1	2	3	4	
		17. The teacher's questions triggered divergent modes of thinking.	0	1	2	3	4	
		18. There was a high proportion of student talk and a significant amount of it occurred between and among students.	0	1	2	3	4	
		19. Student questions and comments often determined the focus and direction of classroom discourse.	0	1	2	3	4	

Student/ Teacher Relationships (20 points)	20. There was a climate of respect for what others had to say.	0	1	2	3	4	
	21. Active participation of students was encouraged and valued.	0	1	2	3	4	
	22. Students were encouraged to generate conjectures, alternative solution strategies, and ways of interpreting evidence.	0	1	2	3	4	
	23. In general the teacher was patient with students.	0	1	2	3	4	
	24. The teacher acted as a resource person, working to support and enhance student investigations.	0	1	2	3	4	
	25. The metaphor "teacher as listener" was very characteristic of this classroom.	0	1	2	3	4	

Table 2. Toshiba Cup score table

Evaluation content	Evaluation index	Value	Score
CONTENT (10 points)	1. Good at grasping the curriculum standards, pay attention to flexible integration of teaching content, adapt to the differences in students' personality and ability training.	3	
	2. The contents of lectures and activities are scientific, ideological and logical.	4	
	3. The teaching focus is prominent, the difficulty and depth are appropriately controlled, and attention should be paid to the connection with the students' existing knowledge and experience.	3	
PROCESS (25 points)	4. Be good at setting questions to guide learning methods, and guide students to actively explore around key and difficult problems.	5	
	5. Pay attention to creating situations in teaching, interact with teachers and students tacitly, and the classroom atmosphere is active and orderly.	5	
	6. The application of teaching methods is reasonable, teaching steps are clear, and students participate effectively.	5	

	7. The teaching is enlightening and vivid, and the explanation is rigorous in logic, clear in thinking, accurate in knowledge, and able to draw inferences from one another.	4	
	8. Flexible handling of teaching events, reflecting teaching wisdom.	3	
	9. According to the feedback information, the teaching process and difficulty can be adjusted appropriately, and the rhythm of teaching process can be controlled in a certain degree.	3	
SKILL (25 points)	10. Teaching demonstration (or experimental demonstration) is standard and proficient.	5	
	11. The design of blackboard writing, board drawing and courseware is reasonable, scientific and beautiful.	5	
	12. Questions are enlightening, and the problem analysis is accurate and comprehensive.	5	
	13. Use mandarin, the language is vivid and clear, the expression is accurate, concise and easy to understand, and the speaking speed is appropriate.	5	
	14. Can effective control time, flexibly use the skills of organizing classroom activities.	5	
INNVATION (15 points)	15. Content innovation: unique situation creation, unique understanding of teaching content, etc.	5	
	16. Means innovation: the design of experimental means is effective, and the design of teaching aids and multimedia courseware are creative, etc.	5	
	17. Form innovation: teaching with characteristics, student activities with characteristics, strong interaction, appropriate guidance of learning methods, etc.	5	
EFFECT (10 points)	18. Teaching objectives are basically achieved.	5	
	19. Promoting the all-round development of students.	5	
COMPREHENSIVE PERFORMANCE (15 points)	20. Dress neatly and decently, behave naturally and generously, have self-confidence and strong affinity.	3	
	21. High level of scientific and humanistic literacy, reflecting discipline thought.	3	
	22. The thinking is agile, flexible, logical, original and critical.	3	

	23. Good at implementing new ideas and methods of curriculum reform in classroom teaching.	3	
	24. Professionalism, caring for students, paying attention to student growth, teaching and cultivating.	3	
Total score:			

2.3. Study 1: Quantitative comparison of RTOP and Toshiba Cup score table

2.3.1. Evaluation object

The evaluation object is a teaching video of teacher Z from H High School on the "First Chinese Middle School Physics Masters Competition". The competition was organized by the Physics Teaching Professional Committee of the Chinese Education Society. Mr. Z won the first prize in this competition. His teaching content was "Newton's Third Law", and the subjects were 32 senior students in C Middle School.

2.3.2. Evaluation subject

The evaluation subjects are divided into two parts. The first part consists of 32 graduate students of physics education in C Normal University as the main part of classroom evaluation, as the basis of other evaluation. They have systematically studied physics, pedagogy and physics classroom teaching evaluation related knowledge, and have strong physics classroom teaching evaluation ability, which lays the foundation for the accuracy of evaluation. The second part is teacher Z, who is a 28 year old male teacher with 6 years teaching experience. Through watching the classroom teaching video, teacher Z also used RTOP and Toshiba Cup score table to evaluate his classroom teaching effect, as the basis for self-evaluation.

2.3.3. Evaluation implementation process

A. Evaluation tool training

The researchers distribute the "Reformed Teaching Observation Protocol (RTOP) Training Manual". The evaluators read the "training manual" by themselves, and put forward questions during the reading process, and the researchers answer them.

The RTOP training manual is comprised of three parts: "structure and core content", "item introduction", and "scoring criteria and cases". The evaluators and teacher Z have 20 minutes to read and understand it. After that, the researchers give a 10-minute answer to any questions that the evaluation subjects put forward during the reading process.

B. Evaluation process

The researchers play the classroom teaching video (length of 41 minutes and 6 seconds), and the evaluators record important classroom events in the blank spaces of the evaluation form while watching the video. After the classroom teaching video is over, the evaluators and teacher Z immediately add notes and finish grading.

The evaluators and teacher Z need to complete the Toshiba Cup score table within 10 minutes after completing the RTOP scale.

2.4. Study 2: Qualitative comparison between RTOP and Toshiba Cup score table

2.4.1. Interview with evaluators

After recalling the RTOP and Toshiba Cup score table, in order to further understand the evaluators' understanding of the advantages and disadvantages of the RTOP and Toshiba Cup score table and suggestions for improvement of this lesson, we randomly select 8 evaluators from 32 evaluators for one-on-one interviews.

2.4.2. Interview with the instructor

After completing data statistics and evaluators interviews, we mainly conduct in-depth interview with teacher Z in terms of "understanding of RTOP" and "discussion and improvement of evaluation suggestions".

3. Results

3.1. Results of Study 1

3.1.1. Statistical analysis of score of Toshiba Cup score table

In the Toshiba Cup score table, the average score given by 32 evaluators is very similar to that given by teacher Z, as shown in Table 3. This proves that the evaluators selected in this study have a high theoretical foundation of classroom teaching evaluation, and can diagnose the classroom accurately with the Chinese classroom teaching evaluation scale. In terms of the total score, the score given by the evaluators and the teacher are more than or equal to 80, which is relatively high in China's classroom teaching evaluation. Therefore, this is consistent with the class winning the first prize in the competition.

Table 3 .Scoring of each subscale by different evaluation subjects using Toshiba Cup score table

Subscale	Average score of evaluators	Standard deviation	Self-evaluation score
Subscale 1: content	8.06	1.045	8
Subscale 2: process	21.19	2.235	20
Subscale 3: skill	21.41	2.108	19
Subscale 4: innovation	11.47	2.016	12
Subscale 5: effect	8.41	1.043	8
Subscale 6: comprehensive performance	12.19	2.162	13
Total score	82.72	10.609	80

3.1.2. Statistical analysis of RTOP score

The score of RTOP is shown in Table 4. In terms of the total score, the average score given by 32 evaluators is 68.38, and that of teacher Z is 64, which are very close. To a certain extent, this reflects that the two evaluation subjects have good consistency in scoring. For each RTOP item, Teacher Z's score is slightly different from the average score of the 32 evaluators. Specifically, there is a big difference in the scores of questions 2, 6, 7 and 14. As for the reasons for the big difference in the scores of these four questions, we have analyzed them in the previous paper, and we will not repeat them here.

Table 4. RTOP scores of different evaluation subjects (Tong, Xing, & Zheng, 2020)

Item	Evaluators		Self-evaluation score	Item	Evaluators		Self-evaluation score
	Average score	Standard deviation			Average score	Standard deviation	
1	3.03	0.801	3	14	1.75	1.047	3
2	3.22	0.751	1	15	2.50	1.136	3
3	2.13	1.264	2	16	2.78	1.070	3
4	2.28	0.813	3	17	2.81	0.965	3
5	1.91	0.963	1	18	2.09	0.928	2
6	3.66	0.545	2	19	1.87	0.942	2
7	3.38	0.660	2	20	3.09	0.963	3
8	3.44	0.504	3	21	3.50	0.622	3
9	2.84	0.847	3	22	2.50	0.950	3
10	3.41	0.712	3	23	3.16	0.987	3
11	2.75	1.016	2	24	2.69	1.120	3
12	2.56	1.045	3	25	2.13	1.008	2
13	2.88	0.907	3	Total score	68.38	12.608	64

3.1.3. Comparison of RTOP and Toshiba cup score

Compared with table 3 and table 4, it can be seen that both the evaluators and the teacher can maintain the consistency of scoring whether using the Toshiba Cup score table or RTOP, and the self-evaluation score of teacher Z is always slightly lower than the average score given by the evaluators. We also find that the total score of different classroom evaluation tools is quite different. Specifically, the score obtained by using Toshiba Cup score table is about 15 points higher than that by RTOP, which is a big difference. Then a question arises: is RTOP not suitable for the classroom teaching evaluation environment in China? However, existing researches have solved this question. The evaluation and promotion group of ACEPT evaluated 141 mathematics and science courses with RTOP scale, and established the norm of RTOP score. The RTOP mean and standard deviation of this sample are 51.3 and 20.1 (Piburn, 2000). MacIsaac and Dan have studied the RTOP score of physics. They think that the teaching with RTOP score more than 50 can be regarded as teaching with reform tendency (MacIsaac, & Dan, 2002). Therefore, the 68.38 points and 64

points respectively given by the evaluators and the instructor in our research should be regarded as relatively high scores in the existing norms of RTOP. In short, the teaching video that won the first prize in the Chinese teaching skills competition can get relatively high scores using the Toshiba Cup score table and RTOP, but the RTOP scoring requirements are more stringent.

3.2. Results of Study 2

3.2.1. Analysis of the results of the evaluator interviews

Through the interviews, it is found that all the 8 evaluators interviewed were exposed to the RTOP scale for the first time. It can be seen that although this scale was introduced to China as early as 2011 (Wu, 2011), it did not have a great impact. As for the difference between RTOP and Toshiba Cup score table, most evaluators believe that RTOP pays more attention to students' learning and procedural knowledge evaluation, while Toshiba cup score table pays more attention to teachers' teaching. When it comes to the advantages of RTOP, the evaluators generally think that RTOP is comprehensive, specific, scientific and highly operable. For example, some evaluators think that "seeing it (RTOP), even if you don't see the classroom teaching video, you can also get a professional direction guide. Even if you can't evaluate and have not learned the evaluation of this kind of teaching, you can carefully consider some of the teacher's teaching design through its guidance." When asked which evaluation form is more in line with the philosophy of China's new curriculum reform, the evaluators unanimously chose RTOP. It can better reflect the autonomy of students, more conducive to the cultivation of students' core literacy. Therefore, in the context of the new curriculum reform, RTOP has a high acceptance and good application prospect.

As for the deficiency and improvement of RTOP, the evaluators brought up mainly two points. First of all, in terms of the scoring, the evaluators think that the RTOP scoring rules are too strict. If the RTOP is used for classroom evaluation, it is difficult for teachers to get a higher score. In addition, the corresponding item characteristics of each score should be further clarified. Secondly, in terms of evaluation content, evaluators believe that RTOP places too much emphasis on the evaluation of students, while ignoring the evaluation of teachers. This is incompatible with China's traditional evaluation of how to teach. In order to better adapt to China's classroom teaching evaluation environment, some items of teacher evaluation should be added. As for the teaching improvement suggestions, the evaluators generally put forward the following three points. First, the teacher should give students more opportunities to ask questions; second, the teacher should leave enough time for students to answer questions, and should be more targeted when asking questions; third, the teacher should make clear the difference between balance force and interaction force.

3.2.2. Analysis of the results of the teacher interview

In the process of interview with teacher Z, the content mainly involves two aspects: asking teacher Z's views on RTOP and suggestions for improvement, and discussing the common problems raised by 32 evaluators with teacher Z on teaching improvement.

First of all, teacher Z and the evaluators have some consensus on the understanding of RTOP. Like the evaluators, Mr. Z is also exposed to the RTOP scale for the first time and believes that the scale is more in line with the concept of the new curriculum reform. Teacher Z thinks RTOP is more operational. He said, "After reading the introduction of RTOP, I probably know how to teach this class. It is very instructive for teaching. Not only in the teaching skills competition, we refer to RTOP before the usual teaching, and we will do better in this class." In addition, Mr. Z believes that RTOP focuses on the development of students' divergent thinking and describes the phenomenon in various ways. These evaluation indicators are quite consistent with the trend of the college entrance examination in China in recent years, which focuses on

"one question with multiple solutions". As for the deficiency of RTOP, Mr. Z says that in Chinese classroom culture, it is not realistic to let students' discussion opinions determine the focus and direction of classroom; due to the lack of students' inquiry ability, it is difficult for students to carry out independent inquiry before class. Teacher Z thinks that RTOP should increase the score of reflection item, and the explanation of the item should be refined to enhance its guidance.

Then, in terms of teaching improvement, teacher Z accepts the suggestions put forward by the evaluators. First, the teacher did not give students too many opportunities to ask questions. Mr. Z responds that the classroom atmosphere was relatively relaxed during the trial lectures in his school, and the students asked many unexpected questions. However, in the formal competition, due to the unfamiliarity between the teacher and students and the more restrictive atmosphere of the competition class, the classroom was not so open. In fact, this is also directly related to the four "grinding lessons" of Mr. Z before the competition. Through the trial lectures, he has mastered the questions that students may raise comprehensively. In order to successfully complete the competition class, he guides students to ask questions that are conducive to the progress of teaching, while avoiding some problems that deviate from the center of the classroom. Second, as for the fact that the teacher did not leave enough time for students to think in class, and no students answered the question after the teacher asked the question, the evaluators suggested that the teacher should leave enough time for the students to answer the question, and should be more targeted when asking the question. Teacher Z agrees very much, he says, "If I look back now, the whole rhythm is a little tight. And really, as the evaluators say, I should give the student some more time to think about it, to answer it, and it should be so." "By asking more specific questions, they may be able to understand more easily and know a direction of inquiry." Third, Teacher Z admits that he did not make the difference between balance force and interaction force clear, which caused certain difficulties to students' understanding. However, he believes that due to the large amount of teaching content, the difference between balance force and interaction force will generally be discussed in the second class.

3.2.3. Comparison of features between RTOP and Toshiba Cup score table

Based on the interviews with 8 evaluators and the instructor, it can be concluded that the RTOP and Toshiba Cup score table have the following characteristics. First of all, they have different focuses on the evaluation. RTOP pays more attention to the learning of students and the evaluation of procedural knowledge; while the Toshiba Cup score table pays more attention to the teaching of teachers. Secondly, they differ in terms of evaluation guidance. RTOP provides more detailed instructional materials such as item introduction and scoring standards, so users feel strong operability. The Toshiba Cup score table does not have similar guiding materials. This is not that we deliberately did not provide these materials, but in China's classroom teaching evaluation environment, evaluation tools are generally used by classroom teaching evaluation experts. By default, they are familiar with these operation specifications and processes, so there are almost no similar guidance materials in the development of evaluation tools. Finally, they differ in the evaluation concept. Both evaluators and the teacher think that RTOP is more in line with the concept of China's new curriculum reform. Because it can better reflect the students' autonomy, more conducive to the cultivation of students' core literacy. However, the evaluation concept reflected in the Toshiba Cup score table is relatively backward in the era of core literacy.

4. Summary and Discussion

We tested the applicability of RTOP in the Chinese classroom teaching evaluation environment from both qualitative and quantitative perspectives, and the effect was good. In this research, the quantitative study of Study 1 shows that using the RTOP and Toshiba Cup scoring tables to evaluate the same excellent class can get higher scores. For the 32 evaluators, their average scores were 68.38 and 82.72 respectively

using these two scales. For the teacher, he used these two scales for self-evaluation and got 64 and 80 respectively. Then, in the qualitative study of Study 2, by comparing the advantages and disadvantages of RTOP and the Toshiba Cup score table, we find that RTOP has the advantages of focusing on students' learning and strong operability, which is more in line with the concept of China's new curriculum reform. In addition, through the use of RTOP, the evaluators proposed the deficiencies of this class, and Mr. Z accepted these deficiencies and put forward specific measures for the improvement of classroom teaching. Therefore, localized research on RTOP can provide valuable reference for the development of classroom teaching evaluation tools in the context of China's new curriculum reform.

As far as physics education is concerned, in January 2018, China's *"The General Senior Secondary School Physics Curriculum Standards (the 2017 version)"* was released, which defined the core quality of physics into four aspects: "physical concept", "scientific thinking", "scientific inquiry", "scientific attitude and responsibility". (Ministry of Education. 2018) Then, in September 2019, the new high school physics textbooks were put into use. As the concept of core literacy is put into practice step by step in the discipline of physics, the next problem is: what kind of physics classroom should be based on the core literacy of physics, and how to evaluate it? The real realization of the development of students' physics core literacy depends on the reform of physics classroom to a great extent, and the classroom teaching evaluation table has a strong regulatory effect on the classroom reform. Therefore, the choice of evaluation tools is very important. However, there is little research on the field of Chinese physics education, which is still in the exploratory stage.

By analyzing the 25 items of RTOP, it is found that it covers the requirements of the core literacy of physics, and has a strong guidance for the reform of Chinese physics classroom. Specifically, "content" contains the requirements of "physical concept". For example, "the lesson involved fundamental concepts of the subject", "connections with other content disciplines and/or real world phenomena were explored and valued", and "intellectual rigor, constructive criticism, and the challenging of ideas were valued" all meet the development needs of students' "physical concept". Secondly, "lesson design and implementation" includes the requirements of "scientific inquiry". For example, items such as "the lesson was designed to engage students as members of a learning community" and "this lesson encouraged students to seek and value alternative modes of investigation or of problem solving" provide good evaluation indicators for "scientific inquiry". In addition, the requirements for "scientific thinking" and "scientific attitude and responsibility" are scattered among the other items in the RTOP. For example, questions such as "students were encouraged to generate conjectures, alternative solution strategies, and ways of interpreting evidence" and "the teacher's questions triggered divergent modes of thinking" are the evaluation indicators for the "scientific thinking" dimension. For instance, "student exploration preceded formal presentation" and "there was a climate of respect for what others had to say" can be used as evaluation indicators of "scientific attitude and responsibility".

There are two deficiencies in this study. First, due to the limitation of specialty and evaluation time, we only study one physics course, so the number of research object is not enough. In order to fully verify the adaptability of RTOP in China, it is still necessary to test multiple courses and disciplines. Second, only one scoring case is given in the RTOP training manual, which is insufficient for the evaluator to understand the scoring rules. In future research, more training time should be left and more scoring cases should be added to reduce measurement errors.

References

- Adamson, S. L., Banks, D., Burtch, M., Cox, F., Judson, E., Turley, J. B., & Lawson, A. E. (2003). Reformed undergraduate instruction and its subsequent impact on secondary school teaching practice and student achievement. *Journal of Research in Science Teaching*, 40(10), 939–957.
- Addy, T. M. , & Blanchard, M. R. . (2010). The problem with reform from the bottom up: instructional practises and teacher beliefs of graduate teaching assistants following a reform-minded university teacher certificate programme. *International Journal of Science Education*, 32(8), 1045-1071.
- Amrein-Beardsley, A. , & Popp, S. E. O. . (2012). Peer observations among faculty in a college of education: investigating the summative and formative uses of the reformed teaching observation protocol (RTOP). *Educational Assessment Evaluation & Accountability*, 24(1), 5-24.
- Campbell, & Todd. (2012). Scientific inquiry in the genetics laboratory: biologists and university science teacher educators collaborating to increase engagement in science processes. *Journal of College Science Teaching*, 41(3), 74-81.
- Erdogan, I. , Campbell, T. , & Abd-Hamid, Nor Hashidah. (2011). The student actions coding sheet (SACS): an instrument for illuminating the shifts toward student entered science classrooms. *International Journal of Science Education*, 33(10), 1313-1336.
- Kulm, G. E. , & Malcom, S. M. E. . (1991). Science Assessment in the Service of Reform. *Washington, DC: American Association for the Advancement of Science*.
- Liu, Z., & Xu, B. (2018). Forty Years Research on Classroom Teaching Evaluation in China: Review and Prospect. *Curriculum. Teaching material. Teaching method*, 38(07), 12-20.
- MacIsaac, & Dan. (2002). Reforming physics instruction via RTOP. *Physics Teacher*, 40(8), 479-485.
- Ministry of Education. (2018). The General Senior Secondary School Physics Curriculum Standards (the 2017 version) (In Chinese). *People's Education Press: Beijing*.
- Pan, S. (2012). Teaching Skills Competition: An Important Way for Science Teachers' Professional Development (in Chinese). *Research on Modern Basic Education*, 8(4), 30-35.
- Piburn, M. D.. (2000). Reformed teaching observation protocol (RTOP) reference manual. technical report. *Classroom Observation Techniques*, 44.
- Sawada, D. , Piburn, M. D. , Judson, E. , Turley, J. , & Bloom, I. . (2010). Measuring reform practices in science and mathematics classrooms: the reformed teaching observation protocol. *School Science & Mathematics*, 102(6), 245-253.
- Tong, D., Xing, H., & Zheng, C. (2020). Localization of Classroom Teaching Evaluation Tool RTOP — — Taking the Evaluation of Physics Classroom Teaching as An Example (in Chinese). *Educational Science Research*, (11), 31-36.
- Wu, W. (2011). An Introduction to RTOP as a Specialized Tool of Classroom Teaching Evaluation (in Chinese). *Research on Teacher Education*, 23(005), 76-80.
- Xing, H., & Tian, W. (2020). The Classroom Teaching Evaluation Theory: Reflection and Construction. *Curriculum. Teaching material. Teaching method*, 40(06), 53-58.