Interactive computer assessment and analysis of students' ability in scientific modeling

Jing Lin [1*] (iD), Letong Zhang [1] (iD), Wenting Wei [1] (iD), Ping-Han Cheng [2] (iD), Chun-Yen Chang [3,4] (iD)

[1] Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University, Beijing, CHINA
[2] Department of Science Education, National Taipei University of Education, Taipei City, TAIWAN
[3] Science Education Center, Graduate Institute of Science Education and the Department of Earth Sciences, National Taiwan Normal University, Taipei City, TAIWAN
[4] Department of Biology, Universitas Negeri Malang, INDONESIA

Abstract

Scientific modeling (SM) is a core scientific practice and critical for students' scientific literacy. Previous research has not used interactive computer assessment to investigate students' SM ability. This study aimed to explore an effective way in human-computer interaction to reveal the challenges faced by students in the four-element process of constructing, using, evaluating, and revising models. Contextualized in the solar system, eleven interactive tasks assessed 419 students in grades 4, 7, and 10. Results indicated that "model evaluation" and "model revision" were more difficult for students than "model construction" and "model use." Grade significantly predicted students' SM ability (p<.001). The interaction with re-answer according to feedback promoted students' in-depth reflection and performance in SM. Findings of the study may provide a basis for improving students' SM ability.

Keywords: interactive computer assessment, scientific modeling, interactive task

INTRODUCTION

Since information technology (IT) penetrates all aspects of our everyday lives nowadays, IT's revolutionary influence naturally also affects educational assessment. Program for International Student Assessment (PISA) (OECD, 2019), Trends in International Mathematics and Science Study (TIMSS) (Mullis & Martin, 2017), and National Assessment of Educational Progress (NAEP) (NAGB, 2019) have implemented the human-computer interaction (HCI)-based method, also known as interactive computer assessment (ICA), which is indicative of the international reform in educational assessment (Farrell & Rushby, 2016; Zainuddin et al., 2020). Compared with paper-and-pencil tests used for decades, ICA is still relatively new. However, ICA did not appear spontaneously but results from many empirical studies in the past half a century. Studies have shown ICA's unique value in assessing higher-order thinking and abilities, and even attitudes (e.g., Baker & O'Neil, 2002; Kuo et al., 2019; Seifried et

al., 2020). ICA has advantages over traditional modes of testing with respect to economy, safety, and convenience in implementation (e.g., Ko & Cheng, 2008; Nissen et al., 2018; Terzis et al., 2012). In the past 30 years, international education assessment reform has highlighted the interaction between artificial intelligence and the human brain in ICA tasks (Burkhardt & Pead, 2003; Silva et al., 2021; Thelwall, 2000; Thurlow et al., 2010). HCI in ICA enables a more intelligent assessment of student literacy to promote the quality education in the 21st century more than ever before.

Scientific modeling (SM) is a scientific practice that involves higher-order thinking and is often viewed as a core component of scientific literacy (Schwarz et al., 2009). Science has been described as ''the process of constructing predictive and conceptual models for predicting or explaining relevant phenomena,'' which means that science is essentially a SM process (Gilbert & Justi, 2016). In this view, to improve students' scientific literacy, science education and assessment should focus on students' SM ability (Schwarz et al., 2022).

✉ linjing@bnu.edu.cn **(*Correspondence)** ✉ 17860717906@163.com ✉ 578948472@qq.com ✉ phcheng@mail.ntue.edu.tw
✉ chunyenc@gmail.com

Contribution to the literature

- This study developed and validated an effective ICA tool diagnosing students' performance during the four-element process of scientific modeling.
- Model evaluation and model revision were more challenging for students than model use and model construction.
- Only grade significantly predicted student modeling ability, while student interest in and time devoted to science learning did not, revealing a need for change in Chinese science classrooms.

Accordingly, the PISA framework for science assessment has the specific requirement of identifying, using, and generating explanatory models and representations (OECD, 2019, p. 108). The TIMSS science assessment requires student to collect data, present and organize relevant data in various intuitive ways, and explore the relationship between variables in the data (Mullis & Martin, 2017, p. 55). NAEP requires explaining relevant scientific principles, connecting different principles, and identifying theoretical models' data types (NAGB, 2019). China's science curriculum standards also include requirements regarding students' SM ability. For example, the *Physics Curriculum Standards* for high schools emphasize model-based teaching to contribute to students' key competencies of scientific thinking (MOE, 2017). China's national science education quality monitoring used a paper-and-pencil test in 2017 to assess students' ability to understand and use scientific models for explaining phenomena and solving problems. However, the number of items assessing SM was small and students' performance was poor (MOE, 2018).

Previous research has indicated that providing students with visual representations can facilitate students' SM (Barak & Hussein-Farraj, 2013; Chang, 2022; Chittleborough & Treagust, 2008). The computer-based interactive tasks can present visual representations of phenomena and provide timely feedback to students' performance, which can create a supportive SM environment for students. This study used an ICA tool around SM with Chinese students in compulsory education to investigate and promote the students' SM ability through a data-driven approach. Furthermore, previous studies have indicated that students' SM ability develops with maturity (Fortus et al., 2016; Pierson et al., 2017; Schwarz et al., 2012), but the relationship between other individual factors and SM is largely unclear to date. It is worthwhile to examine individual differences from the four-element process of SM. Hence, the research questions for this study were:

1. How do the grades 4, 7, and 10 students perform in the process of SM?
2. Does the SM performance of students vary by individual differences in gender, grade, time devoted to and interest in science learning?
3. Is the timely feedback in ICA tool effective in promoting student modeling?

THEORETICAL PERSPECTIVES FOR SCIENTIFIC MODELING ASSESSMENT

Cognitive Domains and Methods for Scientific Modeling Assessment

Cognitive processes of scientific modeling

SM is broadly referring to the practice of constructing and using models to explain or predict corresponding phenomena (NRC, 2012). Models are a systematic abstraction, simplification, and characterization of a phenomenon (Justi & Gilbert, 2002). Judging from preexisting research, SM is complicated and challenging for students (Dori & Kaberman, 2012; Ruppert et al., 2019). Halloun (1996) divided students' SM process into five elements: the selection, construction, validation, analysis, and deployment of models. Students must first select a model framework according to their purpose of use, then reorganize the components and structure of the selected model to construct their own models, then use a variety of methods to assess these models, then use the models to analyze data collected about the phenomena or problems, and finally use the models to explain the phenomena or solve the problems.

Schwarz et al. (2009) further refined the SM process of students from the perspective of cognitive processing as a ''four-element'' process of construction, usage, evaluation, and revision of a scientific model. Schwarz et al. (2009) defined the first element as model construction, which integrates the model selection and construction of Halloun's (1996) theory. Schwarz et al. (2009) deemed that students built an initial model according to their prior knowledge. Students may refer to an existing similar model or not. They may need to detect the related components in the phenomenon and define the structure to build a new model. Second, students need to use their initial models to explain or predict the relevant phenomena, i.e., model use. Students should clarify the structure and function of the model they have constructed through practices such as phenomenon interpretation and problem-solving. Third, students evaluate their own models based on their explanatory or predictive power, i.e., model evaluation. Students should apply the model to a new problem to identify the strengths and weaknesses of their own models. Fourth, students consider various evidence or new phenomena to revise their models according to the findings of model

evaluation, i.e., model revision. Students should summarize and reflect on their models to further refine and improve their modeling products. In other words, students start from their own knowledge to develop their own models of a phenomenon through preliminary construction, usage, evaluation, and final revision.

This "four-element" process is consistent with the concept of interaction learning based on practice proposed by evidence-based practice and design-based practice (Cohen, 2011). The four elements also support high-quality instruction. Model-based instruction is not a manual activity for students or the studying of teacher-made models. Only by constructing, using, and correcting their own models can students develop their understanding of related phenomena (Peel et al., 2019; Zangori et al., 2017). For example, when students evaluate and revise models, they should consider the components which characterize the phenomenon's important characteristics, the relationship between these components, and the rules governing these relationships (Krell et al., 2015). Therefore, the theory of the four elements of SM is pedagogically sound and practical.

Schwarz et al. (2009, 2012) assessed students' SM from two dimensions, generative and dynamic. The generative dimension describes "the reflective practice for how models predict or explain aspects of phenomena when models are constructed or used," which takes SM as a generative tool for predicting or explaining the phenomena. The dynamic dimension describes "reflective practice for when and how models need to change when students evaluate and revise them," which reflects SM is related to an individual's understanding of scientific knowledge and the nature of science. Schwarz et al. (2012) create four sub-dimensions that fit both the generative and dynamic dimensions. The first sub-dimension is "attention to the model's level of abstraction," which describes models constructed by students as literal depictions of, or abstract generalizations of phenomena. The second is "attention to audience and communication clarity," which describes how well students attend their audience's knowledge of the phenomena being modeled and the components that comprise their models. The third is "attention to evidence or authority," which describes the nature of the evidence or support used in students' models. The last sub-dimension is the "nature of the relationship between model and phenomena," which describes how well the model could explain and predict the mechanism and process of phenomena.

Drawing on the generative and dynamic dimensions of Schwarz et al. (2009, 2012), and follow-up studies that presented some more detailed sub-dimensions (e.g., Fortus et al., 2016; Pierson et al., 2017), this study took the four-element theory as the process of SM and also referred to the two dimensions of Schwarz et al. to assess students' SM ability.

Assessment methods for scientific modeling

Preexisting research into students' model-based learning is mainly based on cognitive information processing theory (Gagné, 1975) and has used interviews (e.g., Grosslight et al., 1991; Justi & Gilbert, 2002, 2003), questionnaires (e.g., Chang & Chiu, 2009; Fortus et al., 2016), and classroom observations (e.g., Ke & Schwarz, 2021; Pierson et al., 2017). In early studies, some researchers (e.g., Chang & Chiu, 2009; Halloun, 1996; Lin & Chiu, 2008) used questionnaires to assess students' SM ability and their disciplinary knowledge. Based on the study of Halloun (1996), Chang and Chiu (2009) constructed a modeling structure with 6 elements including model selection, model construction, model validation, model analysis, model deployment, and model reconstruction. The researchers used this structure to test grade 10 students' SM ability and their knowledge of the "galvanic cell," then analyzed and described students' performance level.

To assess grade 8 students' SM process and products, Pierson et al. (2017) used a combination of interviews, paper-and-pencil tests, classroom observations, and performance-based assessments. The study followed the students for a semester, during which students focused on the questions:

(1) why is the soil in a garden moister in some places than in others?
(2) how do roots interact with the surrounding environment?
(3) how can plants help our communities?

During the course, students learnt in the classroom, participated in the measurement of data in a garden, and used the data to construct, evaluate, and revise their own models. The longitudinal data collected by the researchers included video recordings of classroom, audio recordings of the researcher's conversations with students, students' written reflections on SM, students' SM products, and the researcher's field notes during the classroom observation. Students were also interviewed and tested at the end of the course.

In the study by Pierson et al. (2017), students constructed multiple different models, including diagrammatic models which reflected the relationship between roots and the environment, physical models which represented the physiological activities of plants, and simulation-based computational models developed by runnable simulation, which simulated the growth of roots. Pierson et al. (2017) indicated that students' computational models reflected most of their SM ability.

To summarize, most studies assessed students' knowledge of models rather than their SM practices and ability. Some studies have indicated that context is important in the assessment of SM ability and that the computational models are useful for assessing SM ability. Therefore, this study used computers to simulate specific contexts and designed interactive tasks to test

students' SM ability to engage in the four-element process of SM.

Key Points for Designing Interactive Computer Assessment

Unlike paper-and-pencil tests, ICA provides two-way feedback: it can not only provide users feedback on their performance, but also ''learn'' from the users' responses and make changes accordingly (Carlson, 1994). Hence, the purpose of ICA is not to simply digitize paper-and-pencil tests, nor to replace performance-based assessments in real situations, but to make full use of HCI's advantages to assess students' higher-order thinking and practices more effectively (Kuo et al., 2019; Zainuddin et al., 2020). To do so, the ICA tool needs to match the assessment content and its representations with users' cognition to reflect their abilities.

Task analysis

A challenge for ICA is developing the interactive tasks, which includes to determine the indicators and requirements for assessment and ''task analysis'' (Crystal & Ellington, 2004) to design the script for the ICA software development. Three main task analysis methods have been used to date: technical methods, conceptual methods, and work process methods.

Technical methods take a behavior-based perspective (Annett & Duncan, 1967; Kadir & Broberg, 2021) and use hierarchical task analysis (HTA). HTA will analyze and represent tasks' behavioral aspects (Fyiaz et al., 2018), then categorize tasks, break them into subtasks, and check the tasks' overall accuracy.

Conceptual methods focus on analyzing the ''black box'' of cognition (Mason et al., 2019). One of the main techniques is cognitive modeling analysis (Norman, 2008), which creates valuable insights into ''natural mappings'' between cognition and interface. Another technique is the model human processor (Card et al., 1983; Kitajima & Toyota, 2012), consisting of three interacting systems (i.e., perceptual, motor, and cognitive), which aims to map out the restrictions imposed on behavior by the nature and features of the task environment and to define what and when users know about the task. Cognitive task analysis (Hoffman & Militello, 2012) targets more abstract, advanced cognitive functions of the tasks, which requires some subject-matter experts to deeply engage in the particular knowledge domain to analyze various tasks (Chipman et al., 2000).

Work process methods take the perspective of ''activity'' to do task analysis (Bedny & Meister, 1999; Panteli & Kirschen, 2015). Activities are inherently context-sensitive (Nieuwenhuis et al., 2005; Otto & Vassena, 2021). The technique of activity analysis focuses on the entire activity process in a specific context, which needs to balance the activity's efficiency and effectiveness and establish the most economically optimized context through empirical research (Hashim & Jones, 2014).

This study integrated the perspectives of the above three main methods for task analysis. Technical methods' perspective was taken in the hierarchical decomposition of ICA tasks based on the cognitive process of SM. Conceptual methods' perspective was adopted to design the matching between the tasks and the participants' cognitive interaction according to the participants' existing cognitive level. Work process methods' perspective was put in the arrangement of the overall task flow according to the four-element process of SM.

Usability test

Interactive task development should fully consider the users' features such as age, experience, interest to make tasks useful, easy, and fun (Nikou & Economides, 2019). If the task interface is complicated and confusing, the users are prone to negative psychological effects, and their responses will be affected (Guler et al., 2014). Both boys and girls are willing to engage the test if the contents are useful, interesting, and clear (Terzis et al., 2012). The attractive colors, simple and clear words, and readable tasks were employed to build the friendly interface to engage students in the ICA in this study.

The ''usability test'' (Albert et al., 2010) is conducted during the process of ICA task development in this study. Usability testing is a ''user-centric'' technical approach (Preece et al., 2002; Stragier et al., 2013). By observing, recording, analyzing, and judging the users' responses, usability tests detect the users' effectiveness, efficiency, and satisfaction to finish the tasks. The effectiveness indicates the tasks' reliability and validity. Efficiency analysis is to assess whether the resources used by the ICA software are optimized to promote users to achieve their performance. Satisfaction analysis observes the users' experience and feelings to eliminate discomfort, tension, anxiety, and other interfering factors. The ''think aloud'' method (Padilla & Leighton, 2017) in one-on-one interview was taken in this study to let users express their feelings about the test, their thoughts on problem-solving, and their evaluation for tasks, etc. The ICA software was modified according to data analysis on usability test results. Then another round of usability test started again until a satisfactory software product was produced.

Influencing factors

During the task development process, this study also focused on controlling these factors such as gender, computer familiarity, computer self-efficacy, computer-based test expectations, computer anxiety, and computer hardware features which affect ICA assessments (Bahar & Asil, 2018; Skryabin et al., 2015). Boys are more

Table 1. The construct of the solar system scientific modeling: summarized definitions and relevant items

Indicators	Requirements	Modeling the Solar System	Item (score)
Model construction	Constructing a model based on prior knowledge about the planets in the solar system	Students construct a 2D model to represent the planets' positions in the solar system	1 (8)
		Students construct a 2D model to represent the planets' sizes in the solar system	2 (8)
Model use	Using the newly constructed models to explain relevant phenomena	Students explain the brightness of each planet observed from the Earth	3 (8)
		Students explain the brightness of each planet with its distance from the Sun	4 (8)
		Students explain the size of different planets	9 (8)
Model evaluation	Based on new information, evaluating whether the model can explain relevant phenomena	Students evaluate their own models based on the feedback of their responses	5 (8)
		Students further evaluate their models based on new information provided	7 (8)
Model revision	Synthesizing new information and revising the model to explain the phenomena better	Students combine the brightness with distance information of each planet to revise their own models	6 (8)
		Students further consider the different types of planets to revise their models	8 (8)
		Students revise their own models based on the feedback of their responses	10 (8)
		Students reflect on their responses in the previous steps to further revise their models	11 (9)

attracted to ICA if the assessment has practicality. For girls, the simpler the operation system, the more they can be motivated to invest in ICA (Terzis & Economides, 2011). The participants of this study had experienced in computer operation and computer-based assessments, because the first experience with ICA will affect students' expectations of success, which in turn influences the effort students invest in the assessment process (Hewson & Charlton, 2019; Meyer et al., 2016; Timmers & Veldkamp, 2011). Students with different computer self-efficacy levels show significantly different ICA performance (Hewson & Charlton, 2019; Liaw & Huang, 2012). Computer anxiety can interact with test anxiety, leading to poorer performance (Lu et al., 2016; Norris et al., 2007). For the same ICA assessment, the higher the image resolution, the better the students' performance (Guler et al., 2014; Wang et al., 2018). Hence, the images were preferred in designing the task interfaces, the operation requirements were simple with dragging the mouse to answer questions in this study.

METHOD

Participants

Participants were recruited in the study via a convenience sampling. First, a call for volunteer schools was issued to partner schools of the first corresponding author's university, elaborating the objective of the test and the requirements for participating schools and students. There were 21 schools responded initially. After further confirmation of the computer facilities and available date, 10 schools were selected from five Chinese provinces of Heilongjiang, Hunan, Shandong, Shanxi, and Zhejiang. All participating schools had

classroom equipped with internet accessible computers and their students had experienced computer-based assessment. The teachers who were specifically responsible for the test received an instruction manual and online training. In addition, each school was required to double check the network and computers the day before the test and confirmed that there were no problems with the facilities.

Second, another call for research volunteers was delivered to students at participating schools. Regarding of the time and labor costs, we only released the call to 4th-, 7th-, and 10th-grade students who represented the primary, middle, and high school students respectively, to explore the grade differences in student SM performance. A total of 603 students volunteered to take the test. However, some schools were still disconnected during the test, or the interface was too slow to load, certain students did not submit their answers successfully. We finally received responses from 462 students. After excluding vacancies or repeated submissions, 419 valid student-response were obtained, with a reliability of Cronbach's α=.732. There were 168 students in grade 4, 199 in grade 7, and 52 in grade 10; 230 were girls, and 189 were boys.

Construct of Scientific Modeling Assessment

Combined with specific scenarios and tasks, the construct of ICA task for assessing SM is illustrated above (see **Table 1**). The solar system was employed as the context for the test because the galaxies and planets are a type of macroscopic models in space science. It has been indicated that SM ability is crucial for understanding space science phenomena and concepts (Plummer et al., 2016, 2022; Sung & Oh, 2018), which

means modeling the solar system is valid to test students' SM ability. In China, space science is not one of the high school entrance examination disciplines, and as its presence in the university entrance examination is also very low. As a result, most students know little about space science in general and the solar system in particular. Thus, modeling the solar system can be a SM task for students with reliability, rather than rotting memory. Furthermore, students were informed of the brightness of the planets with different spatial positions and sizes in the way of computer prompts. With these prompts, students' tasks in the assessment were to construct, use, evaluate and revise corresponding models towards phenomena.

As shown in **Table 1**, the interactive tasks scaffolded students' SM practice. Students first constructed their own initial solar system model based on their prior knowledge (i.e., items 1 and 2 assess students' model construction). Then, students used their models to explain a relevant phenomenon (i.e., items 3, 4, and 9 assess model use). Next, students used new information to reflect on and evaluate their own models (i.e., items 5 and 7 assess model evaluation). Finally, students built off this new information to revise their models (i.e., items 6, 8, 10, and 11 assess model revision). Students were required to consider eight planets comprehensively to respond to the series of tasks. For each planet with right position and association in the solar system in each task, students got one point. The full score of each item was 8 from items 1 to 10, while 9 for item 11 was due to a new (virtual) planet was added to the scene.

Interactive Computer Assessment Tasks Design

The ICA development process of this study is shown in **Figure 1**. The first step was task analysis. The assessment was divided into four major parts with a total of 11 items (see **Table 1**). Students were asked first to consider the planets' far and near positions to construct the model and gradually combine various planetary distance and brightness information. The students then used the model to predict and interpret newly discovered planets in the solar system, including considering the planets' sizes to evaluate and revise the model.

In the second step, to promote students' comfort with the assessment, the interface was designed to be simple, the text was concise and clear, the pictures were soft in color, the feedback was positive and gentle, and there

Figure 1. ICA tasks development process (Source: Authors' own elaboration)

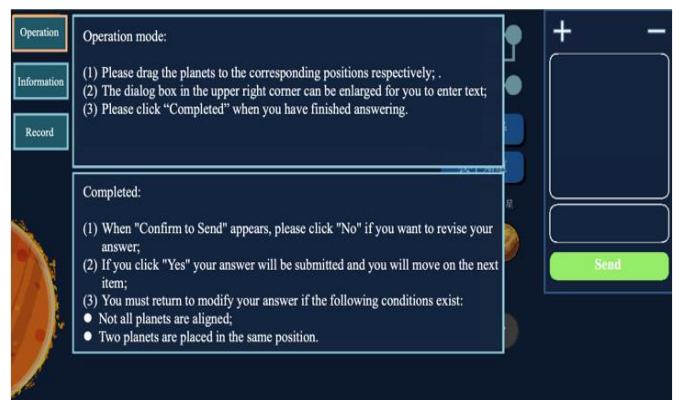**Figure 2.** An interface guiding students to read the tasks and prompts (Source: Authors' own elaboration)

Figure 3. The description of "operation" (Source: Authors' own elaboration)

were question-and-answer prompts to avoid students' anxiety (see **Figure 2** and **Figure 3** for the test guide interface and operation description).

Except for the two model evaluation items, which required students to type text, all other items only required students to move and click a computer mouse. In this way we minimized the influence of the students' computer familiarity on their performance. The computer provided students with information in three ways: one was providing task information to support students' modeling process; second was providing timely feedback for the students to reflect upon when they used the model; the third was to provide an opportunity to re-answer to promote students' reflection and revision during the SM practice.

There were three rounds of the usability tests. After the task script was formed, two students in grade 4, two students in grade 7, and two students in grade 10 were tested individually and interviewed by the researchers to revise the items' assessment points and textual expressions. After converting the script into ICA software, the second round of tests and interviews was conducted, this time focusing on students' thoughts on the interface, answering experience, problem understanding, problem-solving ideas, etc. When the software was developed, a pilot test (α=.701) was done

Table 2. Mean (SD) of the four SM elements

	Model construction	Model use	Model evaluation	Model revision
4th grade	2.76 (2.63)	4.46 (2.82)	0.70 (0.93)	3.50 (3.17)
7th grade	4.19 (2.92)	5.12 (2.95)	1.21 (1.44)	3.69 (3.21)
10th grade	7.10 (2.09)	6.33 (2.63)	2.78 (2.22)	3.97 (3.25)

with a selection of students from grades 4, 7, and 10 using a think aloud approach. Further revisions were made to the tasks based on the pilot test data.

Items for Interest in Science Learning

Interest in science learning was measured with six four-point Likert-type items to exam students' attitude towards science and science learning (Cronbach's α=.881). Two of the items assessed attitudes towards science, and four were about making judgments on their own science learning. For example, students were demanded to self-report their attitude towards "I think science is interesting" or their judgement towards "Learning science is important to me because it always stimulates my thinking" on a scale from 1="totally disagree" to 4="totally agree."

Data Collection

The ICA tool was deployed with automatic recording. Participants answered in their classrooms equipped with computers for 40 minutes. In addition to the 11 items, there were three sections which inquired into students' basic information (school, gender, grade), the hours they devoted to science learning every day and their interest in science learning. IT teachers in charge of the test opened each computer to present the task interface before the test. During the test, the teachers asked students to raise their hands to indicate that they had finished answering all items and waited for teachers to check their submission before leaving their seats.

Data Analysis

Data was first summarized by means and standard deviations (SDs) for the four-element scores of SM. Then, multiple linear regression was used to test the dependency of the four-element scores of SM on gender, grade, science learning time, and science learning interest. Shapiro-Wilk test was conducted to test the normality of scores obtained before and after feedback. Wilcoxon rank-sum test was used to test the effect of the feedback information of ICA. All analyses were conducted with SPSS 25.0.

FINDINGS

Performance of Students in Interactive Computer Assessment

Table 2 indicates that the mean scores of students in "model construction" and "model use" were higher than those of "model evaluation" and "model revision" and
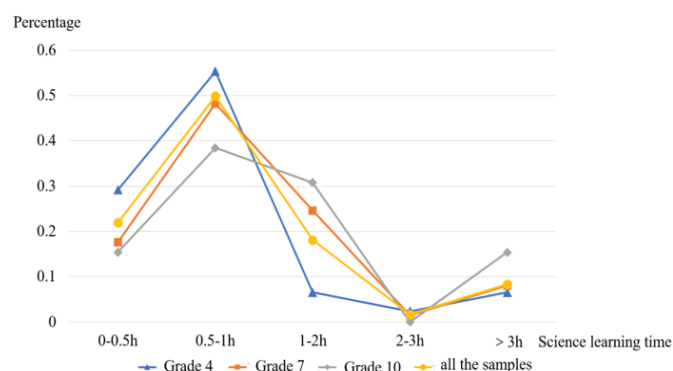

Figure 4. Comparison of students' learning time for science per day (Source: Authors' own elaboration)
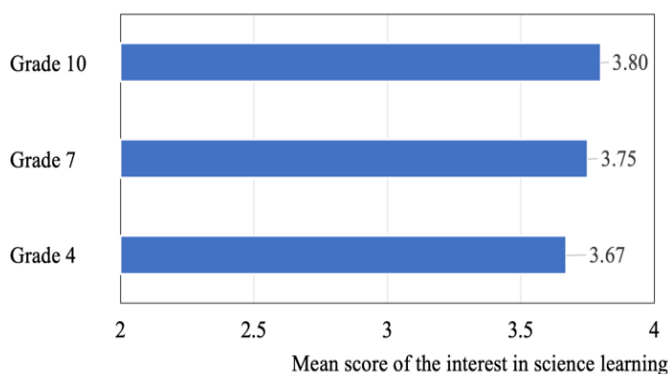

Figure 5. Comparison of interest in science learning for the three grades (Source: Authors' own elaboration)

that the mean score of "model evaluation" was clearly the lowest, for all grades.

Figure 4 and **Figure 5** show students' time for and interest in science learning. It could be seen from that the number of students in grade 10 who studied science for 1~2 hours and more than 3 hours per day was relatively high, and the attitude of thinking that science and learning science were important was also relatively strong than other graders.

Analysis of Related Factors of Scientific Modeling

The results of the correlation analysis showed that students' grade, science learning time, and science learning interest were significantly and positively correlated with their SM ability (p<.05), while gender was not significantly correlated with SM ability (p=.804).

The results of the multiple linear regression analysis are shown in **Table 3**. The Durbin-Watson value was 1.605, between 1.5 and 2.5, indicating this study's sample was strongly independent. All VIF values in each dimension were less than five, verifying the absence of collinearity between the variables. The adjusted R² was

Table 3. The analysis of related factors of SM ability

	Sβ	Sig.	VIF
Constant		.043	
Grade	.419	.000***	1.057
Gender	.007	.869	1.056
Science learning time	.061	.182	1.086
Science learning interest	.080	.073	1.034

Note. Sβ: Standardized β; Sig.: Significance; & ***p<.001

Table 4. The outcomes of the Wilcoxon rank-sum test

Order	Mean	C	Z	Sig.
1st	3.92	.731	-19.231	.000***
2nd	5.38			

Note. Sig.: Significance; C: Correlation; & ***p<.001

.204, indicating that the four independent variables in the study explained 20.4% of the students' SM ability; there were likely other important factors that were not addressed in this study. Among the four independent variables, only the grade significantly predicted the SM ability (p<.001).

One-way analysis of variance (ANOVA) was used to further test the difference among three grades. F-test result (p=.613>.05) examined the homogeneity of variance, indicating ANOVA was appropriate. ANOVA results were significant (F=54.344; df=2, 416; p<.001) and a post-comparison using LSD indicated that all three groups were significantly different from the others (p<.001). The SM ability of the senior students was significantly higher than that of the junior students which was significantly higher than that of the primary students.

The Effect of Feedback Information

The option of "answer again" in the test allowed students to improve their models based on the feedback that the system gave to them. Students' second answers on the items that had a re-answer option (items 3, 4, 6, 8, 9, and 10) were compared to the original responses to determine whether students improved their performance when answering again.

The total number of re-answered answers was 930. A Shapiro-Wilk normality test was performed on the data, indicating that the data was non-normally distributed (p<.001). Therefore, the initial and repeated response results were compared using the Wilcoxon rank-sum test (see **Table 4**). A significant difference between the two responses was identified (p<.001). When students answered the second time, after receiving feedback, their models improved.

CONCLUSIONS AND DISCUSSION

The originality of the study lies in developing an ICA to assess students' SM ability based on the four-element process of SM through task analysis and usability test approaches. The results show that the most challenging task in the process of SM is model evaluation, followed by model revision. Considering that the test in this study demands students to enter text in response to the tasks of model evaluation, which may have increased the item difficulty over the multiple-choice tasks in the other three elements, the conclusion drawn forth on the SM

challenge for students is that model evaluation and model revision are more difficult than model construction and model use. Previous studies have also found that grade 10 students have poor performance in evaluating a model (Chang & Chiu, 2009) and very few grade 6 students are able to revise models (Bamberger & Davis, 2013).

The results of this study also show that individual maturity contributes to students' SM ability, but not gender, science learning time nor science learning interest. These results seem to indicate that spending more time learning science does not appear to improve students' SM ability. One possibility might be that students have fewer learning opportunities for SM. The report for China's national science education quality monitoring in 2017 tells that 4th and 8th grade students have insufficient scientific practice opportunities and poor performance in SM (MOE, 2018). The results of the study also seem to indicate that like or dislike of science plays little role in SM ability. But the finding is possible if students have relatively few opportunities for scientific practices in their science learning which may result in students disconnecting SM from science. The findings of the study may be an indication that what is being taught in the classrooms needs to change.

Previous studies have found that reflection is crucial for SM (e.g., Fortus et al., 2016; Pierson et al., 2017; Schwarz et al., 2009, 2012). The results of this study indicate that the interaction with re-answer according to feedback can promote students' in-depth reflection and performance in SM. The findings from this study can motivate science educators to take more advantage of HCI to create immediate feedback and re-answer opportunities to advance students' learning, and thus promote the cultivation of scientific practices for each student in response to the mission at this historical moment.

The study has the following several limitations. The first is that the study does not collect the data and analyze the influence of individual computer self-efficacy, computer anxiety, and other factors on the test from the physiopsychological perspective, so it is impossible to know whether these factors affect the SM ability between different grades. Second, many school computers cannot be used for the test due to the browser's requirements to get the test interface, limiting the samples in this study. Third, the item type is relatively single. In the four-element process of SM, multiple-choice questions can be added to the model evaluation part, and the model use part can be designed

with a combination of ''prediction-observation-interpretation'' type items.

This present study is the initial research of the SM ICA. Further studies need to focus on the following two aspects. One is to integrate multi-technical methods to collect multi-modal data to explore the complicated SM process. For example, the computerized wearable psychophysiological measurement equipment such as a heart rate variability monitoring and feedback system, can be used to record the changes in the difference in the heartbeat cycle of students to assist in the analysis of the cognitive challenges encountered by the students in the SM process. Non-invasive brain imaging techniques, such as functional near-infrared spectroscopy, can also collect students' cerebral cortex responses and combine the ICA test results to analyze students' cognitive load during the four-element process. In addition, the automated scoring system and teamwork ICA for SM can also be studied in the future. All in all, looking forward to more research on the in-depth development of quality education with smarter assessment to promote the cultivation of scientific literacy for each student.

REFERENCES

Albert, B., Tullis, T., & Tedesco, D. (2010). *Beyond the usability lab: Conducting large-scale online user experience studies*. Morgan Kaufmann Publishers. https://doi.org/10.1016/B978-0-12-374892-8.00009-0

Annett, J., & Duncan, K. D. (1967). Task analysis and training design. *Journal of Occupational Psychology, 41*, 211-221. https://files.eric.ed.gov/fulltext/ED019566.pdf

Bahar, M., & Asil, M. (2018). Attitude towards e-assessment: Influence of gender, computer usage and level of education. *Open Learning: The Journal of Open, Distance and e-Learning, 33*(3), 221-237. https://doi.org/10.1080/02680513.2018.1503529

Baker, E. L., & O'Neil, H. F. (2002). Measuring problem solving in computer environments: Current and future states. *Computers in Human Behavior, 18*(6), 609-622. https://doi.org/10.1016/S0747-5632(02)00019-5

Bamberger, Y., & Davis, E. (2013). Middle-school science students' scientific modelling performances across content areas and within a learning progression. *International Journal of Science Education, 35*(2), 213-238. https://doi.org/10.1080/09500693.2011.624133

Barak, M., & Hussein-Farraj, R. (2013). Integrating model-based learning and animations for enhancing students' understanding of proteins structure and function. *Research in Science Education, 43*, 619-636. http://doi.org/10.1007/s11165-012-9280-7

Bedny, G., & Meister, D. (1999). Theory of activity and situation awareness. *International Journal of Cognitive Ergonomics, 3*(1), 63-72. https://doi.org/10.1207/s15327566ijce0301_5

Burkhardt, H., & Pead, D. (2003). Computer-based assessment: A platform for better tests? In C. Richardson (Ed.), *Whither assessment* (pp. 133-148). Qualifications and Curriculum Authority. https://www.mathshell.com/papers/pdf/hb_dp_2002_whither.pdf

Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction*. Erlbaum. https://doi.org/10.1201/9780203736166

Carlson, R. D. (1994). Computer adaptive testing: A shift in the educational paradigm. *Journal of Educational Technology Systems, 22*(3), 213-224. https://doi.org/10.2190/QP36-WV9L-ATT3-FYWK

Chang, C., & Chiu, M. (2009). The development and application of modeling ability analytic index: Take electrochemistry as an example. *Chinese Journal of Science Education (Taiwan), 17*(4), 319-342. https://doi.org/10.6173/CJSE.2009.1704.04

Chang, H.-Y. (2022). Science teachers' and students' metavisualization in scientific modeling. *Science Education, 106*, 448-475. https://doi.org/10.1002/sce.21693

Chipman, S. E., Schraagen, J. M., & Shalin, V. L. (2000). Introduction to cognitive task analysis. *Cognitive Task Analysis, 1*, 1-8. https://corescholar.libraries.wright.edu/psychology/465

Chittleborough, G., & Treagust, D. (2008). Correct interpretation of chemical diagrams requires transforming from one level of representation to another. *Research in Science Education, 38*, 463-482. https://doi.org/10.1007/s11165-007-9059-4

Cohen, B. J. (2011). Design-based practice: A new perspective for social work. *Social Work, 56*(4), 337-346. http://doi.org/10.1093/sw/56.4.337

Crystal, A., & Ellington, B. (2004). Task analysis and human-computer interaction: Approaches, techniques, and levels of analysis. *Americas Conference on Information Systems, 391*, 1-9. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1967&context=amcis2004

Dori, Y. J., & Kaberman, Z. (2012). Assessing high school chemistry students' modeling sub-skills in a computerized molecular modeling learning environment. *Instructional Science, 40*, 69-91. https://doi.org/10.1007/s11251-011-9172-7

Farrell, T., & Rushby, N. (2016). Assessment and learning technologies: An overview. *British Journal of Educational Technology, 47*, 106-120. https://doi.org/10.1111/bjet.12348

Fortus, D., Shwartz, Y., & Rosenfeld, S. (2016). High school students' meta-modeling knowledge. *Research in Science Education, 46*, 787-810. https://doi.org/10.1007/s11165-015-9480-z

Fyiaz, K., Tabassum, S., & Hasnain, A. (2018). Enhancement of user experience by hierarchical task analysis for interaction system. In I. Nunes (Ed.), *Advances in human factors and systems interaction* (pp. 427-438). Springer. https://doi.org/10.1007/978-3-319-60366-7_40

Gagné, R. M. (1975). Observing the effects of learning. *Educational Psychologist, 11*(3), 144-157. https://doi.org/10.1080/00461527509529140

Gilbert, J. K., & Justi, R. (2016). *Modeling-based teaching in science education*. Springer. https://doi.org/10.1007/978-3-319-29039-3

Grosslight, L., Unger, C., Jay, E., & Smith, C. L. (1991). Understanding models and their use in science: Conceptions of middle and high school students and experts. *Journal of Research in Science Teaching, 28*(9), 799-822. https://doi.org/10.1002/tea.3660280907

Guler, C., Kilic, E., & Cavus, H. (2014). A comparison of difficulties in instructional design processes: Mobile vs. desktop. *Computers in Human Behavior, 39*, 128-135. https://doi.org/10.1016/j.chb.2014.07.008

Halloun, I. (1996). Schematic modeling for meaningful learning of physics. *Journal of Research in Science Teaching, 33*, 1019-1041. http://doi.org/1019-1041.10.1002/(SICI)1098-2736(199611)33:9<1019::AID-TEA4>3.0.CO;2-I

Hashim, N., & Jones, M. (2014). *Activity theory: A framework for qualitative analysis.* https://ro.uow.edu.au/commpapers/408/

Hewson, C., & Charlton, J. P. (2019). An investigation of the validity of course-based online assessment methods: The role of computer-related attitudes and assessment mode preferences. *Journal of Computer Assisted Learning, 35*, 51-60. https://doi.org/10.1111/jcal.12310

Hoffman, R. R., & Militello, L. G. (2012). *Perspectives on cognitive task analysis: Historical origins and modern communities of practice.* Psychology Press. https://doi.org/10.4324/9780203809877

Justi, R. S., & Gilbert, J. K. (2002). Modeling, teachers' views on the nature of modeling, and implications for the education of modelers. *International Journal of Science Education, 24*(4), 369-387. https://doi.org/10.1080/09500690110110142

Justi, R. S., & Gilbert, J. K. (2003). Teachers' views on the nature of models. *International Journal of Science Education, 25*(11), 1369-1386. https://doi.org/10.1080/0950069032000070324

Kadir, B. A., & Broberg, O. (2021). Human-centered design of work systems in the transition to industry 4.0. *Applied Ergonomics, 92,* 103334. https://doi.org/10.1016/j.apergo.2020.103334

Ke, L., & Schwarz, C. V. (2021). Supporting students' meaningful engagement in scientific modeling through epistemological messages: A case study of contrasting teaching approaches. *Journal of Research in Science Teaching, 58*(3), 335-365. https://doi.org/10.1002/tea.21662

Kitajima, M. & Toyota, M. (2012). Simulating navigation behavior based on the architecture model human processor with real-time constraints (MHP/RT). *Behavior & Information Technology, 31*(1), 41-58. https://doi.org/10.1080/0144929X.2011.602427

Ko, C., & Cheng, C. (2008). Flexible and secure computer-based assessment using a single zip disk. *Computers & Education, 50*(3), 915-926. https://doi.org/10.1016/j.compedu.2006.09.010

Krell, M., Reinisch, B., & Krüger, D. (2015). Analyzing students' understanding of models and modeling referring to the disciplines biology, chemistry, and physics. *Research in Science Education, 45*(3), 367-393. https://doi.org/10.1007/s11165-014-9427-9

Kuo, B., Liao, C., Pai, K., Shih, S., Li, C., & Mok, M. M. C. (2019). Computer-based collaborative problem-solving assessment in Taiwan. *Educational Psychology, 40*(9), 1053-1055. https://doi.org/10.1080/01443410.2018.1549317

Liaw, S., & Huang, H. (2012). Perceived satisfaction, perceived usefulness, and interactive learning environments as predictors to self-regulation in e-learning environments. *Computers & Education,*

60(1), 14-24. https://doi.org/10.1016/j.compedu.2012.07.015

Lin, J.-W., & Chiu, M.-H. (2008). Exploring high school students' knowledge of models and modeling from cognitive methodology. *Science Education Journal (Taiwan), 307*, 9-14. https://doi.org/10.6216/SEM.200804_(307).0002

Lu, H., Hu, Y. P., Gao, J. J., & Kinshuk. (2016). The effects of computer self-efficacy, training satisfaction and test anxiety on attitude and performance in computerized adaptive testing. *Computers & Education, 100*, 45-55. https://doi.org/10.1016/j.compedu.2016.04.012

Mason, B., Rau, M. A., & Nowak, R. (2019). Cognitive task analysis for implicit knowledge about visual representations with similarity learning methods. *Cognitive Science, 43*(9), 37. https://doi.org/10.1111/cogs.12744

Meyer, A. J., Innes, S. I., Stomski, N. J., & Armson, A. J. (2016). Student performance on practical gross anatomy examinations is not affected by assessment modality. *Anatomical Sciences Education, 9*(2), 111-120. https://doi.org/10.1002/ase.1542

MOE. (2017). Gāozhōng wùlǐ kèchéng biāozhǔn [The curriculum standard for high school physics]. *Ministry of Education of the People's Republic of China*. http://www.moe.gov.cn/srcsite/A26/s8001/202006/t20200603_462199.html

MOE. (2018). Zhōngguó yìwù jiàoyù zhìliàng jiāncè bàogào [Report on monitoring of the quality of compulsory education of China]. *Ministry of Education of the People's Republic of China*. http://www.moe.gov.cn/jyb_xwfb/gzdt_gzdt/s5987/201807/t20180724_343663.html

Mullis, I. V. S., & Martin, M. O. (2017). TIMSS 2019 assessment frameworks. *IEA*. https://timss2019.org/wp-content/uploads/frameworks/T19-Assessment-Frameworks.pdf

NAGB. (2019). Science framework for the 2019 national assessment of education progress. *National Assessment Governing Board*. https://www.nagb.gov/content/dam/nagb/en/documents/publications/frameworks/science/2019-science-framework.pdf

Nieuwenhuis, S., Heslenfeld, D., J., von Geusau, N. J. A., Mars, R. B., Holroyd, C. B., & Yeung, N. (2005). Activity in human reward-sensitive brain areas is strongly context dependent. *NeuroImage, 25*(4), 1302-1309. https://doi.org/10.1016/j.neuroimage.2004.12.043

Nikou, S. A., & Economides, A. A. (2019). A comparative study between a computer-based and a mobile-based assessment: Usability and user experience. *Interactive Technology and Smart Education, 16*(4),

381-391. https://doi.org/10.1108/ITSE-01-2019-0003

Nissen, J. M., Jariwala, M., Close, E. W, & Dusen, B. V. (2018). Participation and performance on paper- and computer-based low-stakes assessments. *International Journal of STEM Education, 5*, 21. https://doi.org/10.1186/s40594-018-0117-4

Norman, D. A. (2008). The way I see it–Simplicity is not the answer. *Interactions, 15*(5), 45-46. https://doi.org/10.1145/1390085.1390094

Norris, J. T., Pauli, R. B., & Bray, D. E. (2007). Mood change and computer anxiety: A comparison between computerized and paper measures of negative effects. *Computers in Human Behavior, 23*(6), 2875-2887. https://doi.org/10.1016/j.chb.2006.06.003

NRC. (2012). A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. *The National Academy Press*. https://nap.nationalacademies.org/catalog/13165/a-framework-for-k-12-science-education-practices-crosscutting-concepts

OECD. (2019). PISA 2018 assessment and analytical framework. *Organization for Economic Co-operation and Development*. https://www.oecd-ilibrary.org/education/pisa-2018-assessment-and-analytical-framework_b25efab8-en

Otto, A. R., & Vassena, E. (2021). It's all relative: Reward-induced cognitive control modulation depends on context. *Journal of Experimental Psychology: General, 150*(2), 306-313. https://doi.org/10.1037/xge0000842

Padilla, J. L., & Leighton, J. P. (2017). Cognitive interviewing and think aloud methods. In B. D. Zumbo, & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research; understanding and investigating response processes in validation research* (pp. 211-228). Springer. https://doi.org/10.1007/978-3-319-56129-5_12

Panteli, M., & Kirschen, D. S. (2015). Situation awareness in power systems: Theory, challenges, and applications. *Electric Power Systems Research, 122*, 140-151. https://doi.org/10.1016/j.epsr.2015.01.008

Peel, A., Zangori, L., Friedrichsen, P., Hayes, E., & Sadler, T. D. (2019). Students' model-based explanations about natural selection and antibiotic resistance through socio-scientific issues-based learning. *International Journal of Science Education, 41*(4), 510-532. https://doi.org/10.1080/09500693.2018.1564084

Pierson, A. E., Clark, D. B., & Sherard, M. K. (2017). Learning progressions in context: Tensions and insights from a semester-long middle school

modeling curriculum. *Science Education, 101*(6), 1061-1088. https://doi.org/10.1002/sce.21314

Plummer, J. D., Bower, C. A., & Liben, L. S. (2016). The role of perspective taking in how children connect reference frames when explaining astronomical phenomena. *International Journal of Science Education, 38*(3), 345-365. https://doi.org/10.1080/09500693.2016.1140921

Plummer, J. D., Udomprasert, P., Vaishampayan, A., Sunbury, S., Cho, K., Houghton, H., Johnson, E., Wright, E., Sadler, P. M., & Goodman, A. (2022). Learning to think spatially through curricula that embed spatial training. *Journal of Research in Science Teaching, 59*(7), 1134-1168. https://doi.org/10.1002/tea.21754

Preece, J., Rogers, Y., & Sharp, H. (2002). *Interaction design: Beyond human-computer interaction*. John Wiley & Sons. https://doi.org/10.1145/512526.512528

Ruppert, J., Duncan, R. G., & Chinn, C. A. (2019). Disentangling the role of domain-specific knowledge in student modeling. *Research in Science Education, 49*(3), 921-948. https://doi.org/10.1007/s11165-017-9656-9

Schwarz, C. V., Ke, L., Salgado, M., & Manz, E. (2022). Beyond assessing knowledge about models and modeling: Moving toward expansive, meaningful, and equitable modeling practice. *Journal of Research in Science Teaching, 59*(6), 1086-1096. https://doi.org/10.1002/tea.21770

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Acher, A., Fortus, D., Schwartz, Y., Hug, B., & Krajcik, J. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching, 46*(6), 632-654. https://doi.org/10.1002/tea.20311

Schwarz, C., Reiser, B. J., Acher, A., Kenyon, L., & Fortus, D. (2012). MoDeLS: Challenges in defining a learning progression for scientific modeling. In A. C. Alonzo, & A. W. Gotwals (Eds), *Learning progressions in science: Current challenges and future directions* (pp. 101-137). Sense Publishers. https://doi.org/10.1007/978-94-6091-824-7_6

Seifried, J., Brandt, S., Kögler, K., & Rausch, A. (2020). The computer-based assessment of domain-specific problem-solving competence–A three-step scoring procedure. *Cogent Education, 7*(1), 1719571. https://doi.org/10.1080/2331186X.2020.1719571

Silva, N., Zhang, D., Kulvicius, T., Gail, A., Barreiros, C., Lindstaedt, S., Kraft, M., Bölte, S., Poustka, L., Nielsen-Saines, K., Wörgötter, F., Einspieler, C., & Marschik, P. B. (2021). The future of general movement assessment: The role of computer vision and machine learning–A scoping review. *Research

in Developmental Disabilities, 110*, 103854. https://doi.org/10.1016/j.ridd.2021.103854

Skryabin, M., Zhang, J., Liu, L., & Zhang, D. (2015). How the ICT development level and usage influence student achievement in reading, mathematics, and science? *Computers & Education, 85*, 49-58. https://doi.org/10.1016/j.compedu.2015.02.004

Stragier, J., Derboven, J., Laporte, L., Hauttekeete, L., de Marez, L. (2013). Kilowhat? A multidisciplinary approach on the development of a home energy management system. *Behavior & Information Technology, 32*(11), 1086-1104. https://doi.org/10.1080/0144929X.2012.740638

Sung, J. Y., Oh, P. S. (2018). Sixth grade students' content-specific competencies and challenges in learning the seasons through modeling. *Research in Science Education, 48*, 839-864. https://doi.org/10.1007/s11165-016-9589-8

Terzis, V., & Economides, A. A. (2011). Computer based assessment: gender differences in perceptions and acceptance. *Computers in Human Behavior, 27*(6), 2108-2122. https://doi.org/10.1016/j.chb.2011.06.005

Terzis, V., Moridis, C. N., & Economides, A. A. (2012). How student's personality traits affect computer based assessment acceptance: Integrating BFI with CBAAM. *Computers in Human Behavior, 28*(5), 1985-1996. https://doi.org/10.1016/j.chb.2012.05.019

Thelwall, M. (2000). Computer-based assessment: A versatile educational tool. *Computer & education, 34*, 37-49. https://doi.org/10.1016/S0360-1315(99)00037-8

Thurlow, M., Lazarus, S. S., Albus, D., & Hodgson, J. (2010). Computer-based testing: Practices and considerations. *National Center on Educational Outcomes, University of Minnesota*. https://nceo.umn.edu/docs/onlinepubs/synthesis78/synthesis78.pdf

Timmers, C., & Veldkamp, B. (2011). Three studies are presented on attention paid to feedback provided by a computer-based assessment for learning on information literacy. *Computers & Education, 56*(3), 923-930. https://doi.org/10.1016/j.compedu.2010.11.007

Wang, J.-H., Chang, L.-P., & Chen, S. Y. (2018). Effects of cognitive styles on web-based learning: Desktop computers versus mobile devices. *Journal of Educational Computing Research, 56*(5), 750-769. https://doi.org/10.1177/0735633117727598

Zainuddin, Z., Shujahat, M., Haruna, H., & Chu, S. K. W. (2020). The role of gamified e-quizzes on student learning and engagement: An interactive gamification solution for a formative assessment system. *Computers & Education, 145*, 103729. https://doi.org/10.1016/j.compedu.2019.103729

Zangori, L., Peel, A., Kinslow, A., Friedrichsen, P., & Sadler, T. D. (2017). Student development of model-based reasoning about carbon cycling and climate change in a socio-scientific issues unit. *Journal of Research in Science Teaching, 54*(10), 1249-1273. https://doi.org/10.1002/tea.21404

https://www.ejmste.com