International Journal of
STEM Education

RESEARCH

Open Access

Measuring science self-efficacy with a focus on the perceived competence dimension: using mixed methods to develop an instrument and explore changes through cross-sectional and longitudinal analyses in high school

Xinyang Hu, Yanxia Jiang* and Hualin Bi

Abstract

Background: In many countries and regions, such as the United States, Europe and China, a trend has emerged in which students' enthusiasm for STEM is declining. This decline may be related to students' lack of science self-efficacy. An accurate examination of students' science self-efficacy can provide a research foundation for how to cultivate it. This paper used mixed methods to develop a valid science self-efficacy scale for high school students, focusing on the perceived competence dimension. A cross-sectional analysis exploring and interpreting differences across grades and genders in science self-efficacy among high school students was conducted. Subsequently, a 1-year longitudinal study was conducted on the development of science self-efficacy in China.

Results: This study developed a 24-item science self-efficacy instrument based on the Rasch model, and the validity of the instrument was assessed through multiple aspects, including face, content, construct, and predictive validity. This instrument was used to divide students' science self-efficacy into four different levels. A cross-sectional study examining 1564 high school students in 10th–12th grades revealed that students' science self-efficacy exhibited a complex process of decreasing and then increasing by grade. Most girls' science self-efficacy was higher than that of boys for Levels 1 and 4, while for the intermediate levels, i.e., Levels 2 and 3, most boys had higher science self-efficacy than girls. The quantitative and qualitative results of the longitudinal study through a 1-year follow-up of 233 high school students indicated that students' science self-efficacy significantly improved. We revealed inconsistencies between cross-sectional and longitudinal studies of the change in science self-efficacy from 10 to 11th grade.

Conclusions: This study makes many contributions. First, we developed a science self-efficacy measurement instrument for high school students with high reliability and validity based on the Rasch model and characterized four different levels of student science self-efficacy. Second, the gender differences in science self-efficacy and the complex changes among grades were explained from the perspective of science self-efficacy level. Finally, students' science self-efficacy significantly improved in the longitudinal study, which was explained by self-efficacy theory and the Chinese core competency-oriented science curriculum.

Keywords: Science self-efficacy, Rasch analysis, Assessment, Validation, High school

*Correspondence: jiangyanxia@sdnu.edu.cn

College of Chemistry, Engineering and Materials Science, Shandong Normal University, Jinan, Shandong, China

Introduction

Science, technology, engineering, and mathematics (STEM) education has become increasingly valued over the past 20 years (Honey et al., 2014). However, students' enthusiasm for STEM learning has declined dramatically in many countries and regions (Thomas & Watters, 2015). In addition, enrollment and participation in the STEM career path have decreased annually (e.g., Marginson et al., 2013; OECD, 2008). Based on the monitoring of the science learning of 200,000 primary and secondary students, the latest National Compulsory Education Quality Testing: Monitoring Results of Scientific Learning Quality Report, published in 2020, showed that only approximately 20% of the students were willing to engage in science-related careers when they grew up. From the perspective of social cognitive factors, researchers have suggested that the lack of science self-efficacy is one of the important reasons for this lack of willingness (Ballen et al., 2017; Blotnicky et al., 2018; Lamb et al., 2014).

Compared with general self-efficacy, self-efficacy in the field of science and other challenging subjects is more important for students' learning (Mangos & Steele-Johnson, 2001). Science self-efficacy can predict students' science academic achievement (Honicke & Broadbent, 2016; Tuan et al., 2005). Students with a higher sense of science self-efficacy have more confidence in their abilities, a greater willingness to successfully complete science tasks, and a stronger perseverance in completing difficult science tasks (e.g., Baldwin et al., 1999; Britner & Pajares, 2006). In contrast, students with low science self-efficacy are more likely to give up on science tasks. If a student feels that he or she lacks the ability to be competent in science tasks, this belief will lead to aversion to science and low academic achievement (Baldwin et al., 1999).

Research is lacking on high school students' science self-efficacy (e.g., Dalgety & Coll, 2006; Gainor & Lent, 1998). Moreover, most previous studies on science self-efficacy have focused on college students (e.g., Ainscough et al., 2016; Baldwin et al., 1999; Uzuntiryaki & Aydın, 2009). However, compared with that of college students, high school students' science self-efficacy requires greater attention for the following reasons: (1) although self-efficacy develops over one's lifetime, as children enter adulthood, self-efficacy gradually solidifies (Bandura, 1994); (2) high school is the period when individuals have the lowest science self-efficacy (Eccles et al., 1997), and many students may no longer learn science after graduating from high school, and their science self-efficacy may also cease to develop (Larose et al., 2006); and (3) the level of high school science self-efficacy may affect students' choice of college or future career. Byars-Winston et al. (2010) found that students with higher science self-efficacy in high school were more likely to choose a STEM major in college.

The level of students' science self-efficacy is constantly changing and can be improved through appropriate teaching intervention and guidance (e.g., Baldwin et al., 1999; Heggestad & Kanfer, 2005). High school is the key period during which to evaluate and cultivate students' science self-efficacy, and it is necessary to accurately measure and cultivate these students' science self-efficacy to improve their academic achievement and professional interest in science.

The main purpose of this study is to develop the Science Self-Efficacy Scale (SSES) with high reliability and validity using mixed methods. Then, this instrument is used to explore changes in high school students' science self-efficacy through cross-sectional and longitudinal analyses. Specifically, our study seeks to answer the following questions:

1. To what extent is the instrument a valid and reliable measure for evaluating students' science self-efficacy across grades 10–12?
2. How can high school students' science self-efficacy levels be used to explore and explain differences across grades and gender differences in science self-efficacy?
3. How does the science self-efficacy of the same students change from grades 10 to 11?

Relevant literature review

Self-efficacy

Self-efficacy describes people's beliefs in their abilities to perform given tasks (Bandura, 2006). Clarification of the constructs of self-efficacy and self-concept is often required before research on self-efficacy is conducted (Bandura, 1997). Self-efficacy refers to one's belief that he or she can successfully complete a task (Bandura, 1986). Self-concept refers to individuals' beliefs and views about themselves, including their judgment of their self-worth (Pajares & Schunk, 2001). A description such as "I am good at science" is a self-conceptual judgment. The statement "I can explain an experimental phenomenon with scientific principles" reflects a self-efficacy belief because it reflects one's judgment of his or her competency to complete tasks in a specific situation (Aydın & Uzuntiryaki, 2009).

Self-efficacy is shaped by cognitive processing and the integration of four main sources of information: mastery experience, vicarious experiences, social persuasion, and physiological and affective states (Bandura, 1986, 1997). This cognitive processing and integration of information from multiple sources determine an individual's self-efficacy. Mastery experience is seen as the prominent source of self-efficacy because it is related to students' interpretations of their previous performance (Bandura, 1997, 2012; Britner & Pajares, 2006; Kıran & Sungur, 2012).

Science self-efficacy

Science self-efficacy is a person's belief in the ability to successfully complete specific tasks in the field of science (Robnett et al., 2015). Britner (2008) proposed that science self-efficacy is a strong predictor of science academic achievement, regardless of gender. Higher science self-efficacy leads to a higher predicted level of science academic achievement (Mataka & Kowalske, 2015; Tezer & Aşıksoy, 2015). This relationship is of central interest to science educators (Lamb et al., 2014). Science self-efficacy and gender appear to be strongly associated, but previous studies have drawn different conclusions (Huang, 2012; Livinţi et al., 2021; Sezgintürk & Sungur, 2020; Weisgram & Bigler, 2006). Girls have been shown to exhibit higher science self-efficacy and higher science achievement than boys in middle school (e.g., Britner & Pajares, 2001; Pajares et al., 2000). However, other researchers have reported that boys' science self-efficacy is higher than that of girls (e.g., Chan, 2022; Weisgram & Bigler, 2006). Still others have found that science self-efficacy does not differ significantly between boys and girls (Rowe, 1988; Sezgintürk & Sungur, 2020). Thus, science self-efficacy may not be a function of gender (Lips, 1992).

Measurement of science self-efficacy

Self-efficacy is very difficult to measure because it is a latent response variable (Scherbaum et al., 2006). Self-efficacy can be measured only indirectly (Judge, 2009), often in the form of self-report surveys (Cassidy & Eachus, 2002). A growing body of quantitative instruments are being designed to measure different aspects of science self-efficacy (e.g., Baldwin et al., 1999; Lamb et al., 2014; Tezer and Asiksoy 2015; Uzuntiryaki & Aydın, 2009). For example, Lamb et al. (2014) viewed science and technology self-efficacy as a unidimensional structure and developed a scale for it based on the Rasch model. Tezer and Asiksoy (2015) developed the physics self-efficacy scale with two dimensions: learning level and physics problem solving. Uzuntiryaki and Aydın (2009) developed the college chemistry self-efficacy scale, which contains cognitive skills, psychomotor skills and everyday application dimensions. The authors also developed the high school chemistry self-efficacy scale, which contains cognitive skills and chemistry laboratory dimensions. Baldwin et al. (1999) developed the biology self-efficacy scale for nonmajors with three dimensions: biology methods, generalization to other biology/science courses and analyzing data, and the application of biological concepts and skills. Tatar et al. (2009) developed the science and technology self-efficacy scale with three dimensions: confidence in science and technology ability, coping with difficulties in science and technology, and confidence in performing science and technology tasks. Çalişkan et al. (2007) developed a physics self-efficacy scale with five dimensions: solving physics problems, physics laboratory, learning physics, and the application and memorization of physics knowledge.

Bandura (1997) suggested that if the self-efficacy construct in question is context or domain specific, it should be unidimensional. However, researchers have different understandings of the number of dimensions of the science self-efficacy structure, ranging from one to five. This fact indicates that no consensus has been reached as to the structure of science self-efficacy and that a convincing conceptual framework is lacking. Furthermore, some factors or dimensions of the existing research on science self-efficacy overlap. For example, Lamb et al. (2014) stated that science and technology self-efficacy as a construct refers to the ability to apply cognitive skills to the overall tasks involved in the use of computers and application of science, which can include the three dimensions proposed by Tatar et al. (2009). On the other hand, the exploration of the structure of science self-efficacy often uses exploratory or confirmatory factor analyses based on classical test theory. A limitation of this method is that it is unable to distinguish between true multidimensionality and spurious correlations attributable to method effects, which are due to the scale or items properties (Finaulahi et al., 2021; Yan Zi, 2010).

In addition, researchers in the field of self-efficacy research agree that quantitative research needs to be complemented by qualitative research to fully describe the complete process of self-efficacy in narrative form (Pajares, 1996; Schunk, 1991). However, only a few qualitative methods have been used in science self-efficacy research (Dalgety & Coll, 2006).

Conceptual framework

Self-efficacy is defined as one's perception of his or her ability to complete a task or be successful in a particular domain (Bandura, 1977, 1982). Based on Bandura's self-efficacy theory, science self-efficacy is conceptualized as students' perceptions of their competence in science tasks. Bong and Skaalvik (2003) proposed that beliefs in one's competency to perform tasks can be regarded as a substitute for students' self-efficacy. Many researchers have also chosen to represent science self-efficacy based on students' perceived competence beliefs in science tasks (Jansen et al., 2015; Larose, et al., 2006). For example, Jansen et al. (2015) adapted a science self-efficacy test based on real science tasks under the science framework of the Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS).

The present study focuses on measuring students' science self-efficacy based on their perception of their own science task competence, which is the major dimension of science self-efficacy (Bandura, 2006). From the practical or functional point of view, science self-efficacy can be considered a unidimensional structure (Bejar, 1983; Smith, 1996). However, unidimensionality is a relative concept (Andrich, 1988; Glynn, 2012). The main requirement of unidimensionality is that the items work sufficiently together to define a trait (Bejar, 1983; Planinic et al., 2019). For example, Uzuntiryaki and Aydın (2009) recognized that the three factors of chemistry self-efficacy all belong to the perceived self-efficacy dimension. Moreover, the unidimensional structure of science self-efficacy has been confirmed by many studies (e.g., Boone et al., 2011; Lamb et al., 2014).

The Rasch model is a type of single-parameter logistic model of item response theory, which is a preferred approach to developing measurement instruments (Boone & Staver, 2020; Liu, 2010; Lu et al., 2020). The Rasch model can place latent person abilities, item difficulties, and level thresholds on the same scale so that comparisons between them are possible (Liu & McKeough, 2005). The Rasch model is a suitable method for analyzing Likert-like scale items, which can convert the ordinal variables to interval variables, making it possible to analyze students' abilities and item difficulties on an equal-interval linear scale. In addition, person ability and item difficulty are set on a true interval scale from negative infinity to positive infinity, which can avoid ceiling and floor effects by using raw scores (Bond & Fox, 2007). Because of these advantages, we chose the Rasch model to develop a science self-efficacy scale.

Bandura (2006) suggested that one of the important issues to be addressed in self-efficacy measurement is the division of the self-efficacy level thresholds. If a low threshold is chosen, low self-efficacy can also be considered full confidence. Conversely, choosing an excessively high level results in the tested individuals consistently being labeled as lacking efficacy. Moreover, "identification of different levels of the defined construct" is an important step in the development of a scale based on the Rasch model (Boone & Staver, 2020; Liu, 2010).

The perception and experience of difficulty of the relevant task is an important factor affecting self-efficacy (Bandura, 1997). Gist and Mitchell (1992) found that the difficulty of tasks directly affects students' self-efficacy levels. Thus, determining learning tasks at different levels of difficulty is key to defining self-efficacy levels. The measurement of science self-efficacy level should be performed according to the level of science task difficulty, which indicates the level of challenges or obstacles students face to achieve success (Bandura, 2006). In addition, these science tasks should have sufficient difficulty levels to avoid ceiling effects.

The science tasks in large international science tests, such as the TIMSS, National Assessment of Educational Progress (NAEP) and PISA, are highly correlated and have similar task difficulty levels (e.g., Bybee et al., 2009; Kind, 2013). The difficulty of science tasks is mainly affected by cognitive processes. If the cognitive processes required to solve different science tasks are similar, these science tasks can be considered to have the same task difficulty level (Rindermann, 2007). Anderson et al. (2001) revised the Bloom classification of cognitive processes, proposing that cognitive processes are defined by six levels. Ranging from low to high, these levels are remember, understand, apply, analyze, evaluate and create. Webb (1999) proposed four levels of cognitive complexity based on the depth of knowledge, namely, recall and reproduction, skills and concepts, strategic thinking/reasoning, and extended thinking. In sum, both international science testing and cognitive psychology research have reached a partial consensus on the difficulty levels of science tasks. Considering the above level framework, science tasks were divided into four levels, as described in Table 1.

Self-efficacy refers to one's conviction that one can achieve a given level of performance in a specific task (Bandura, 1986). If individuals are not confident in facing simple science tasks, their science self-efficacy level is relatively low. On the other hand, if individuals have confidence in facing the most complex science tasks, their science self-efficacy level is regarded as relatively high. Finally, if individuals are confident in being able to address a simple science task, but not confident in addressing the most complex, their level of science self-efficacy is considered to be intermediate. The difficulty level of the science task can be used as a scale or progression variable to guide the development of science self-efficacy items. As noted in Fig. 1, the least difficult items are at the bottom of the scale, and the most difficult items are at the top of the scale, with other moderately difficult items in the middle of the scale (Boone et al., 2011; Wright & Stone, 1979).

Methods

A mixed methods design was used in this study, which specifically included three parts (Creswell & Plano Clark, 2011; Hesse-Biber et al., 2015). The first part involved the development of an instrument for measuring science self-efficacy in high school. An exploratory sequential

Table 1 Consistency comparisons between the difficulty level of the science task, cognitive process and international science tests

TIMSS 2007: science cognitive domain	**NAEP 2009: science practices**	**PISA 2006: competencies**	**Bloom's cognitive process (revised)**	**Webb's depth of knowledge**	**Science task difficulty**	Lower Difficulty ↕ Higher Difficulty
Knowing	Identifying scientific principles	Identifying scientific issues	Remember	Recall and reproduction	Identifying and remembering	
Applying	Using scientific principles	Explaining phenomena scientifically	Understand Apply	Skills and concepts	Explaining and applying	
Reasoning	Using scientific inquiry	Using scientific evidence	Analyze Evaluate	Strategic thinking	Inquiring and argument	
	Using technological design		Create	Extended thinking	Innovative design	

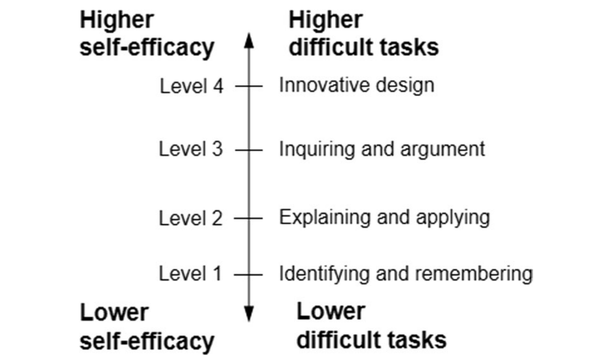


Fig. 1 Theoretical scale of science self-efficacy

design was used for this process, starting with the gathering of qualitative data and resulting in the development of an instrument. The second part was a cross-sectional analysis of science self-efficacy across 10th–12th grades. The third part was a 1-year longitudinal analysis of science self-efficacy from grades 10 to 11 using the explanatory sequential design; that is, qualitative data were used to explain the results of the quantitative research.

Part 1: development of the science self-efficacy scale (SSES)

The validity analysis framework can guide the development of a new instrument (Trochim & Donnelly, 2006). The development process of the science self-efficacy scale (SSES) is shown in Fig. 2.

Stage 1: definition

Defining refers to "using both existing theory and research to provide a sound conceptual foundation to clearly and concretely define the construct being measured"; this is the first step in any scale development (DeVellis, 2017; He et al., 2021). As mentioned above, science self-efficacy is considered unidimensional and can be conceptualized based on students' perceptions of their competency in science tasks. Moreover, the level of science self-efficacy is closely related to the difficulty of science tasks, and the following four levels have been defined: identifying and remembering, explaining and applying, inquiring and argument, and innovative design.

Stage 2: development

This stage refers to the development process of the initial survey, and in the following sections, the main science tasks that high school students might be asked to accomplish are summarized based on interviews with high school students and teachers as well as an analysis of international science tests. Based on these science tasks, the corresponding items of the SSES were compiled. Finally, the face validity and content validity of the initial test instrument were ensured through pairwise comparison and expert group interviews.

A behavior event interview is an objective, open and retrospective competency survey method (McClelland, 1998). To achieve thematic saturation on science tasks

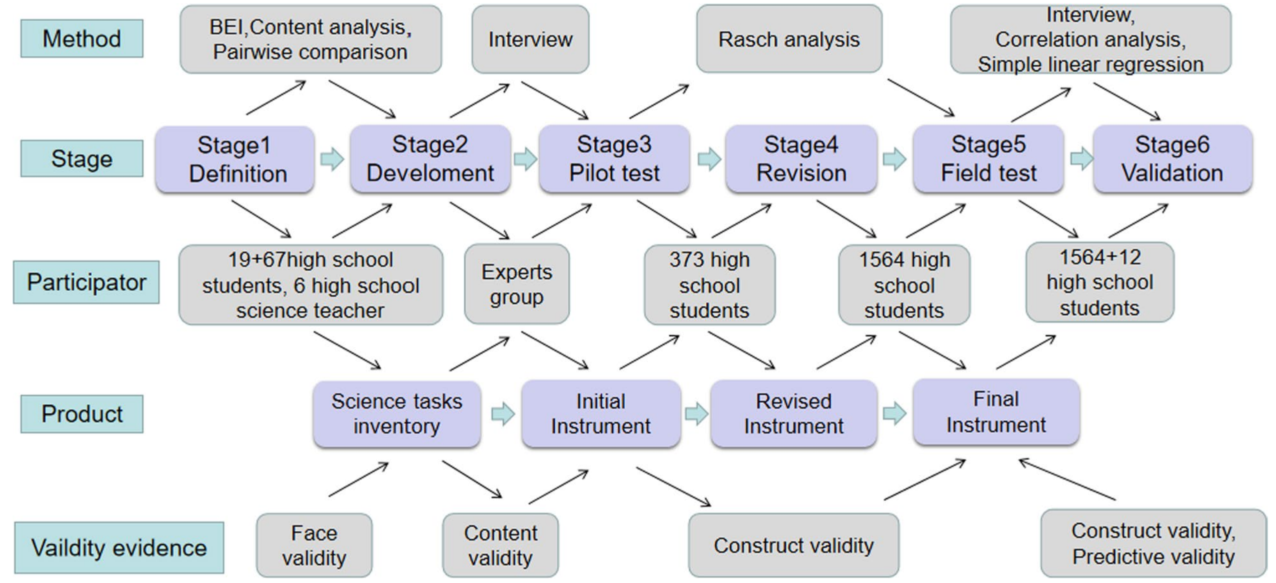


Fig. 2 Development and validation process of the SSES

in the interviews, at least 6–12 participants needed to be interviewed (Johnson & Onwuegbuzie, 2004). Fifteen grade 12 high school students and 4 high school science teachers were interviewed. They were asked to recall and report the science tasks that they were asked to complete (or assigned to their students, in the case of teachers) during high school. To ensure that the participants' descriptions of science tasks were as comprehensive as possible, corresponding science tasks were presented by introducing the science content of the NAEP, TIMSS and PISA and describing previously published test items.

Using deductive content analysis, the coding scheme was confirmed based on the framework of science self-efficacy levels in this study, and the theoretical framework was expanded and validated on the basis of encoding (Cho & Lee, 2014; Neuendorf, 2017). First, student interviews and NAEP, TIMSS, and PISA science tests were identified as the analytical units. Second, the descriptions of the science tasks in the student interviews and the NAEP, TIMSS and PISA science tests were selected as the analytical samples. Third, based on the theoretical framework in Fig. 1, different descriptions of the science tasks in the analytical sample were coded into four levels. The encoding process was performed independently by the two researchers, and the results were recorded using Microsoft Excel spreadsheets. The Cohen kappa of the encoding result was 0.86, which indicated an excellent common-coding protocol (Bakeman & Gottman, 1986). A third researcher was consulted on inconsistent coding results, and a final consensus on coding was reached. Finally, the coding results were integrated and reduced.

Subsequently, four grade 12 high school students and two high school science teachers were interviewed randomly. No new science tasks were reported, which proved that thematic saturation on science tasks had been reached (Francis et al., 2010). The final coding results indicated alignment with theoretical expectations; the 4 levels of science tasks and the corresponding behavioral performance codes as well as examples of sources are given in Table 12 in the "Appendix". Next, 29 items of the initial version of the SSES were designed according to the behavioral performance descriptions of science tasks in Table 13 in the "Appendix". In addition, a 5-point scale ranging from 1 ("I couldn't do this at all") to 5 ("I could do this easily") was used for scoring. Since the subjects of this study were Chinese high school students, the items were all written in Chinese. Then, the scale was translated into English by two researchers who had lived in the United States for more than 1 year. The science task coding process and examples of the item design are as follows:

> In the interview, the students responded to the following item: "Designing and making a simple alcohol detector based on chemical principles and sensors."
> Level: "innovative design"; Code: "Designing and completing production." According to the coding of this science task, item 29 was designed as follows: "Designing and completing a practical technology product independently based on scientific principles."

Face validity

To verify the instruments' face validity, the pairwise comparison method was used to test whether the students' perceived science task difficulty and competency level were consistent with the theoretical assumption of this study (e.g., Oishi et al., 1998).

A total of 67 students from 10th to 12th grades were randomly selected. One science task item was randomly selected from each of the four levels in turn, and the students were given pairwise comparisons of difficulty levels for the science task items. In other words, the subjects were asked to compare the difficulty of science task items at two different levels and to indicate which was more difficult and which was simpler.

Each student needed to complete the following six pairwise comparison questions, which were randomly selected: Level 1–Level 2, Level 2–Level 3, Level 3–Level 4, Level 1–Level 4, Level 1–Level 3, and Level 2–Level 4. Among the students, 51 were able to make correct judgments for all comparisons, 8 students judged correctly in 5 comparisons, 5 students judged correctly in 4 comparisons, and 3 students judged correctly in 3 comparisons. As the subjects of this study, high school students' perceptions of the difficulty of science task items were consistent with the theoretical assumptions, which proved that the science self-efficacy items were effective and meaningful for them; that is, the instrument had face validity (He et al., 2021).

Content validity

To address the content validity, an expert panel of 3 high school science teachers and 4 university science education researchers (2 of whom were also measurement and evaluation experts) was formed. This panel participated in the entire process of scale development and validity testing in this study. A review outline for the expert panel is provided in Table 14 in the "Appendix". Some items were modified according to the experts' recommendations to make them clearer. An example of an item modification is presented below.

Original item: "Finding common characteristics of all kinds of science phenomena or science facts through comparison." According to the experts' recommendations, this item was revised to "Identifying the similarities and differences between different science phenomena."

Stage 3: pilot test

Rasch model-based tests require that subjects have different abilities. Therefore, the cluster sampling method was adopted; that is, the class was taken as the unit to extract the participants. In June 2018, 431 students from 9 classes were selected from grades 10, 11 and 12 of three high schools in Shandong Province, China, for a pilot test. This test was approved by the abovementioned schools, teachers, and students themselves and took 20 min, with the entire test being supervised by the teachers. A total of 431 questionnaires were sent out, and 373 were recovered, for a recovery rate of 86.5%. The choice of this sample size met the basic requirements for the use of the Rasch model for data analysis and processing, i.e., that the number of samples should be more than 200 or 5–10 times the number of test items (Liu, 2010). The data were imported into Winsteps 3.72.0 software for statistical analysis.

To account for potential dependency when calculating data from students within multiple classrooms, grades and schools, the intra-class correlations (ICCs) were calculated (Goldstein, 2011). The ICC of the school level is 0.02 ($p=0.21$), the ICC of the grade level is 0.04 ($p=0.15$) and the ICC of the class level is 0.06 ($p=0.18$), all of which are less than 0.100 ($p>0.05$). These results indicate that the differences among the schools, grades and class levels are not significant and that the potential dependency issue can be ignored (Raudenbush & Bryk, 2002).

According to the results of the pilot test, the person separation index was 2.77 (reliability = 0.88), and the item separation index was 8.35 (reliability = 0.99); both were acceptable (Duncan et al., 2003; Wright & Masters, 1982). A Rasch analysis can provide item fit statistics to show how well each item fits the Rasch expected response structure (Bond & Fox, 2007). Item fit statistics include the mean square residual (MNSQ) and the standardized mean square residual (ZSTD) based on the differences between what was observed and what was expected based on the Rasch model. The MNSQ is a simple squared residual based on the difference between the observed response pattern and the predicted response pattern, and the ZSTD is the normalized *t*-score of the residual. In addition, a Rasch analysis generates infit statistics and outfit statistics, which are the weighted mean square residual and the unweighted mean square residual, respectively. A common standard is that infit and outfit MNSQ values should be within the range of 0.7–1.3 and that the infit and outfit ZSTD values should be within the range of -2.0 to $+2.0$ for acceptable fit (Bond & Fox, 2007). In this pilot study, the infit MNSQs, outfit MNSQs, infit ZSTDs and outfit ZSTDs were all within the acceptable ranges with the exception of a few described below, indicating that the preliminary instrument met the theoretical expectation based on Fig. 1 and had adequate reliability.

Stage 4: revision

According to the item fit statistics of the pilot study sample, it was found that the difficulty levels of items such

Table 2 Item distribution in the SSES levels

Level	Description	Distribution of items
Level 1	Identifying and remembering	Q1; Q2; Q3; Q4; Q5; Q6; Q7
Level 2	Explaining and applying	Q8; Q9; Q10; Q11; Q12; Q13; Q14; Q15
Level 3	Inquiring and argument	Q16; Q17; Q18; Q19; Q20
Level 4	Innovative design	Q21; Q22; Q23; Q24

as Q3, Q11 and Q28 were not completely consistent with the theory. Therefore, these items were adapted descriptively. For example, Q11 was revised from "Using scientific principles to explain daily life" to "Using scientific principles to explain common phenomena in experiments or daily life," which simplified this item. There were some item fit statistics far beyond the acceptable scope. After discussion with the expert group, it was finally decided to delete items Q2, Q6, Q9, Q18 and Q19. The remaining items were unchanged.

After modification of the pilot test items, an instrument with a total of 24 items was formed. The item distribution of the final SSES is shown in Table 2, and the complete final SSES is shown in Table 15 in the "Appendix".

Stage 5: field test

To enable the test sample to cover as many students as possible, 6 high schools in 6 cities in Shandong Province, China, were selected for the test in September 2018. A total of 1903 students from 10 to 12th grades were tested. The number of effective questionnaires was 1564, and the effective rate was 82.2%. The resulting review for this study was obtained from the schools' institutional monitoring board. All participants were informed of the purpose and procedure of the test and voluntarily participated in the test. Both the schools' and students' identities were altered through coding to guarantee their anonymity. Moreover, the students were also informed about and agreed to participate in the follow-up interview research. The whole test process was supervised by teachers, and the test time was 20 min. The demographic information is shown in Table 3.

Table 3 Demographics of the high school students in the field test

Variable	Category	Frequency	Percentage (%)
Gender	Male	698	44.63
	Female	866	55.37
Grades	Grade 10	337	21.55
	Grade 11	655	41.88
	Grade 12	572	36.57
Major	Science	985	62.92
	Nonscience	579	37.08
School type	Urban school	852	54.48
	Rural school	712	45.52
Total		1564	100

Stage 6: validation

To further verify the validity of the scale developed in this study, evidence from cross-validation through triangulation is provided. In addition to quantitative methods, qualitative methods were used to further verify the construct validity of the instrument. In previous studies, interviewees were often asked to describe their self-confidence in science (Webb-Williams, 2018). However, as Bandura (2006) argued, it is difficult for students to judge and describe their level of confidence by themselves. Therefore, specific science tasks were needed to understand the actual science self-efficacy levels of students.

Presenting students with specific science tasks allowed them to articulate their perceptions and judgments of their competence in these science tasks. If a student perceived his or her competence in science tasks to be at or below the level measured by the SSES and perceived himself or herself to be incompetent in higher-level tasks, the student's science self-efficacy level was accurately measured. Based on the division of students' science self-efficacy levels above, 3 students were randomly selected at each level, and 12 students were interviewed voluntarily. Students randomly selected 3 adapted science tasks from each level for 12 science tasks. To avoid the interference of subject knowledge proficiency on students' perceptions of science task competence, the 3 science tasks selected at each level included physics, chemistry and biology. The student's judgment of his or her competence in each level of science task depended on him or her giving two or more of the same answers.

To provide evidence of predictive validity, the science test scores from the students' admissions in September 2018 were collected and converted into *Z* scores, which represent students' science academic achievements (Fiorella et al., 2021; Klein, 2014). Pearson correlation coefficients and a simple linear regression were calculated between students' science self-efficacy and the scores of science academic achievement (Lamb et al., 2014; Pedaste et al., 2021). Moreover, correlation and simple linear regression analyses of science academic achievement were conducted at each level of science self-efficacy separately.

Part 2: cross-sectional analysis

The second purpose of this study was to explore and explain the potential cross-sectional differences in students' science self-efficacy across grades 10–12 in China.

This part involved the application of field test data. Data were collected and analyzed using SPSS 26.0 software, and the dependent variable was the Rasch score (in logits) (Luo et al., 2021). Two-way analysis of variance (ANOVA) was used to explore the differences in science self-efficacy across grade and gender. Then, the exploration and interpretation of the differences across grades and genders in science self-efficacy were conducted by Chi-square analysis using the levels of students' science self-efficacy.

Part 3: longitudinal analysis

Students from five high schools in Shandong Province were selected for the test, and a 1-year longitudinal analysis was conducted. Specifically, the first science self-efficacy survey was conducted at the beginning of grade 10 (September 2019), and 267 students were selected. The second follow-up survey was conducted at the beginning of grade 11 (September 2020), and 233 students, including 119 boys (51.07%) and 114 girls (48.93%), were matched and tracked. Quantitative data were collected and analyzed using SPSS 26.0 software, and the dependent variable was the Rasch score (in logits) (Luo et al., 2021).

Since this study tested the same students' science self-efficacy before and after high school science curriculum learning, a paired-sample *t* test and Chi-square analysis were used to explore how students' science self-efficacy changed from grades 10 to 11. Erickson (2012) suggested that qualitative research is particularly appropriate in education to identify and understand change over time. Therefore, on the basis of collecting and analyzing the above quantitative data, stratified sampling was adopted to interview students from Levels 1 to 4 separately, with 2 students at each level and 1 student below Level 1, for a total of 9 students. These 9 students covered all levels of science self-efficacy and both increased and decreased at each level, thereby meeting the key stratifiers needed for in-depth research in this study; thus, the sample reached saturation (Charmaz, 2006). One year after the SSES was administered, the 9 students were further followed up and interviewed. Since the selection of qualitative interview samples needed to be based on the level of quantitative measurement, the qualitative data were collected only after the quantitative data were analyzed. The combination of quantitative and qualitative data was performed only at the interpretation stage to ensure triangulation of the data sources (Creswell & Plano Clark, 2011).

Results

Part 1: SSES reliability and validity

Unidimensionality and local independence

First, a confirmatory factor analysis was performed using the raw scores obtained from the measurement by Amos 26.0. The results showed that the one-factor model had almost acceptable fit indices (Chi-square = 768.166, $p < 0.000$, DF = 225, Normed Chi-square = 3.414, CFI = 0.910, GFI = 0.959, RMSEA = 0.039, SRMR = 0.025) (Maydeu-Olivares & Joe, 2014; Pedaste et al., 2021).

The unidimensionality of the measurement scale and the local independence of the measurement items are two basic assumptions of the Rasch model (Liu, 2010; Lu et al., 2020). The purpose of the unidimensionality test is to search for data that are inconsistent with the latent trait of science self-efficacy. In other words, it tests whether other dimensions or components affect the students' responses to the items. The unidimensionality of the rating scale was assessed by a principal component analysis of the residuals (PCAR), and a combination of criteria was used to support the claim of the unidimensionality of the measure (Boone & Staver, 2020; Linacre, 2011). If the variance as explained by the Rasch factor was ≥ 20%, it was considered to support unidimensionality. The Rasch model explained 40.1% of the total variance, which was greater than 20%. Furthermore, among these dimensions, the dimension with the greatest influence in the unexplained variance accounted for only 5.7%, which was less than 10% (Boone & Staver, 2020; Linacre, 2011). These findings suggest that although multiple extra dimensions may have influenced students' responses, none of these dimensions had a significant impact. The above suggests that the overall unidimensionality of the measure was acceptable.

Local independence requires that students are not affected by other items when answering a certain item. Item factor loadings within the range of − 0.4 to + 0.4 indicate that those items might measure the same dimension (Liu, 2010). All items were within the range of − 0.4 to + 0.4, except for Q2, Q23 and Q24, with correlation coefficients of 0.44, 0.55 and 0.60, respectively, which required closer examination.

The Rasch model is an ideal mathematical model that is unlikely to be perfectly achieved in realistic measurements because even simple tests may be disrupted by irrelevant factors (Yan Zi, 2010). Therefore, it is difficult for data to perfectly meet the unidimensional and local independence requirements of the Rasch measurement. Deleting certain items requires rigorous consideration of whether these items obviously differ from other items given the assessment aim (Boone & Staver, 2020; Linacre, 2011). Q2 described a very simple science task to evaluate whether students reached Level 1, and Q23 and Q24 described higher-difficulty science tasks aligning with Level 4. Although it is possible that these items were too simplistic or too difficult to have been encountered by students in everyday science learning, they made an important contribution to preventing ceiling effects.

Moreover, the correlation coefficients of all items were less than 0.7, which suggested that the items could be considered highly locally independent (Linacre, 2011; Liu, 2010). After discussion with the expert group, it was finally decided to retain these three items.

Reliability

In Rasch analysis, reliability is a property of the person and the item measured, with two indicators: the person separation index and the item separation index. The separation index can also be converted to Cronbach's α equivalent value, with a range of 0–1. A summary of the statistics of the measurement instrument is presented in Table 4. The revised instrument had a person separation index of 1.97, and the corresponding person reliability was 0.79. In general, a person separation index higher than 2 and a person reliability greater than 0.8 imply that the instrument is sensitive enough to distinguish between high and low performers (Duncan et al., 2003). Although there was still a small gap between the data of this study and the ideal standard, it was acceptable for low-risk assessment (Liu, 2010). The item separation index was 25.85, and the correspondence item reliability was 1.00. In brief, the separation index of the items was very high, and the person separation index was acceptable. In addition, these indices indicated that the spread of items and persons were reliably calibrated along the latent trait measured by the scale.

Validity

Construct validity

Based on the assumption that the SSES can measure the 4 levels of students' science self-efficacy, if the corresponding levels could be divided, it would provide evidence of the structural validity of the instrument developed in this study (Boone & Staver, 2020; Linacre, 2011). Specifically, the process of verifying the structural validity of the instrument according to the Rasch model could be divided into the following three steps (Liu, 2010; Lu & Bi, 2016).

The first step was to ensure that all test items fit the Rasch model. The data fit is typically assessed by infit and outfit mean square (MNSQ). Infit is sensitive to inlier misfit, while outfit is sensitive to outlier misfit (Oon & Fan, 2017). In general, items have acceptable fit if their MNSQ values fall in the range of 0.6 to 1.4 for the rating scale (Linacre, 2011; Liu, 2010). The data shown in Table 5 reveal that the infit MNSQ and outfit MNSQ values of all items were within the acceptable range.

The PTMEA corr. value is the correlation between the person item scores and person measures. For a Rasch analysis, the value should be positive and not be nearly zero (Bond & Fox, 2007). A negative correlation indicates potential fit errors; thus, the higher the positive correlation is, the better (Liu, 2010). Table 5 shows that the PTMEA corr. values for all of the items were all positive and ranged from 0.23 to 0.58, suggesting adequate item discrimination.

Table 5 Fit statistics of the items in the final instrument

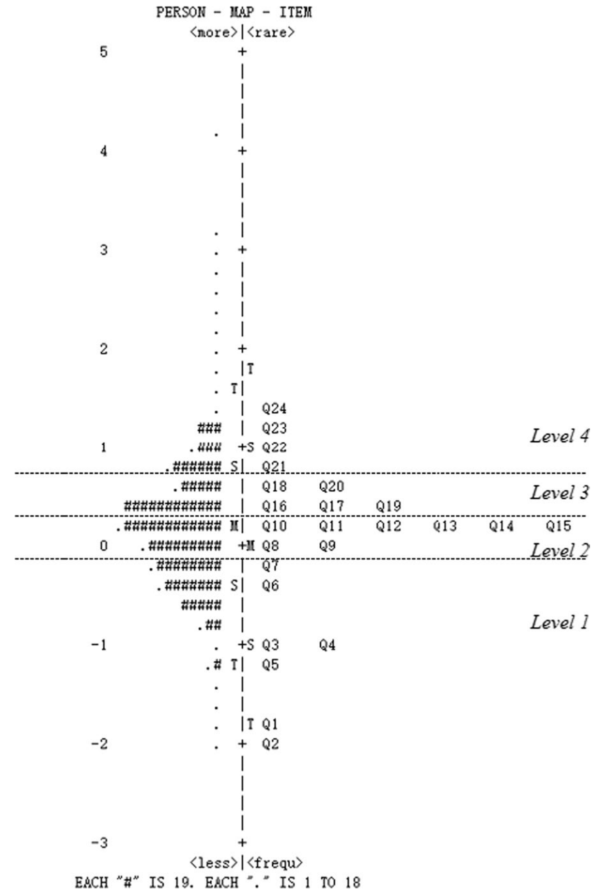
Item	Measure	Model S.E	Infit MNSQ	Outfit MNSQ	PTMEA corr.
Q1	− 2.23	0.04	0.82	0.80	0.25
Q2	− 2.1	0.04	0.85	0.84	0.23
Q3	− 1.1	0.04	0.63	0.62	0.29
Q4	− 1	0.04	0.65	0.65	0.38
Q5	− 1.13	0.04	0.65	0.64	0.40
Q6	− 0.34	0.03	1.07	1.07	0.51
Q7	− 0.29	0.03	1.08	1.07	0.44
Q8	0.01	0.03	1.03	1.04	0.49
Q9	− 0.05	0.03	0.89	0.89	0.58
Q10	0.12	0.03	1.09	1.11	0.53
Q11	0.11	0.03	1.06	1.06	0.51
Q12	0.15	0.03	0.87	0.87	0.53
Q13	0.25	0.03	1.07	1.07	0.52
Q14	0.3	0.03	1.07	1.07	0.51
Q15	0.29	0.03	0.98	0.99	0.55
Q16	0.49	0.03	0.99	1.00	0.52
Q17	0.49	0.03	0.94	0.96	0.53
Q18	0.62	0.03	1.09	1.14	0.42
Q19	0.47	0.03	1.02	1.07	0.46
Q20	0.51	0.03	1.10	1.11	0.46
Q21	0.79	0.03	1.11	1.12	0.35
Q22	0.97	0.03	1.21	1.22	0.40
Q23	1.25	0.03	1.05	1.07	0.35
Q24	1.43	0.03	1.18	1.19	0.26

Table 4 Summary statistics of persons and items

Parameter (*N*)	Measure	Infit		Outfit		Separation	Reliability
		MNSQ	ZSTD	MNSQ	ZSTD		
Person (1564)	0.19	1.01	− 0.3	0.99	− 0.4	1.97	0.79
Item (24)	0.00	0.98	− 0.6	0.99	− 0.3	25.85	1.00

The second step was to verify that the difficulty of the test items was consistent with the level of science self-efficacy constructed in Table 2. In other words, the more difficult the science task item was, the higher the corresponding student's science self-efficacy level. The Wright map (presented in Fig. 3) is a graphical representation of the science self-efficacy levels. In Fig. 3, the left side of the Wright map in the vertical direction is the distribution of the science self-efficacy of the students, indicating low self-efficacy to high self-efficacy from bottom to top. The right side of the Wright map in the vertical direction is the distribution of item difficulty, with the science tasks becoming more difficult from bottom to top.

According to the Wright map, the science self-efficacy level of the students presented a normal distribution, and the item difficulty estimation almost coincided with the students' science self-efficacy level distribution. However, some students' science self-efficacy level still lacked science task items of the corresponding difficulty level. For example, there were still some students with extremely high levels of science self-efficacy who exceeded the difficulty levels of Q23 and Q24, which was beyond our expectations. However, studies have shown that both high achievers and students with learning disabilities often overestimate their abilities (e.g., Klassen, 2002; Pajares & Miller, 1994). This tendency may be difficult to correct by increasing the difficulty level of the science task.



```
               PERSON - MAP - ITEM
                  <more>|<rare>
     5                  +
                        |
                        |
                        |
                        |
                      . |
     4                  +
                        |
                        |
                        |
                        |
                      . |
     3                . +
                      . |
                      . |
                      . |
                      . |
     2                . +
                      . |T
                      . T|
                      . |  Q24
                    ### |  Q23                                   Level 4
     1             .### +S Q22
-----------------.###### S|  Q21--------------------------------------------
                .#####  |  Q18      Q20                          Level 3
           ############ |  Q16      Q17      Q19
----------.############ M|  Q10     Q11      Q12     Q13     Q14     Q15
     0        .######### +M Q8      Q9                           Level 2
--------------.########---|--Q7---------------------------------------------
              .####### S|  Q6
                  ##### |
                    .## |                                        Level 1
    -1                . +S Q3      Q4
                    .# T|  Q5
                      . |
                      . |
                      . |T Q1
    -2                . +  Q2
                        |
                        |
                        |
                        |
    -3                  +
                  <less>|<frequ>
   EACH "#" IS 19. EACH "." IS 1 TO 18
```

Fig. 3 Wright map for the SSES

In addition, there was a gap of approximately 1 logit between Q1 and Q5, which may be due to the large difficulty difference among the tasks in Level 1. For example, some science tasks, such as identifying intuitive information and observing phenomena, only require a one-step cognitive process, while other science tasks, such as identifying specific relationships and accurately describing observation information, are often based on recalling cognitive activities, and they require two or more cognitive processes (Webb, 1999). Overall, these science tasks are very simple. In short, the distribution of the difficulty of most items was consistent with the structure of the theoretical hypothesis in Table 2.

The third step was the calculation of the mean measurements for every science self-efficacy level by averaging the values of all the items at each level. The threshold for students' science self-efficacy was defined as the average of the difficulty values of all items corresponding to each level (Lu & Bi, 2016; Lu et al., 2020). As shown in Table 6, the calculated threshold for Level 1 was − 1.17, the threshold for Level 2 was 0.15, the threshold for Level 3 was 0.52, and the threshold for Level 4 was 1.11. Students' science self-efficacy was divided into four levels, and the specific thresholds are shown in Table 6.

If the measured value of the student's science self-efficacy was less than − 1.17, it meant that the student could not even reach Level 1 and had no perceived competence for most of the science tasks. If a student's test value was between − 1.17 and 0.15, it meant that the student reached Level 1 and was confident only in science tasks involving "identifying and remembering". If a student's test value was between 0.15 and 0.52, this indicated that the student achieved Level 2 and felt competent in "exposing and applying" science tasks but not yet

Table 6 Mean measures of science self-efficacy levels

Level	Item (measurement)	Mean	S.D
1	Q1 (− 2.23); Q2 (− 2.10); Q3 (− 1.10);Q4 (− 1.00); Q5 (− 1.13); Q6 (− 0.34); Q7 (− 0.29)	− 1.17	0.76
2	Q8 (0.01); Q9 (− 0.05); Q10 (0.12);Q11 (0.11); Q12 (0.15); Q13 (0.25); Q14 (0.30); Q15 (0.29)	0.15	0.13
3	Q16 (0.49); Q17 (0.49); Q18 (0.62); Q19 (0.47); Q20 (0.51)	0.52	0.06
4	Q21 (0.79); Q22 (0.97); Q23 (1.25); Q24 (1.43)	1.11	0.29

in "inquiring and argument" science tasks. If a student's test value was between 0.52 and 1.11, it meant that the student achieved Level 3 and felt competent in "inquiring and argument" science tasks but not yet competent in "innovative design" science tasks. If the student's test value was greater than 1.11, it meant that the student reached Level 4 and felt competent in "innovative design" science tasks. Table 6 also shows that the mean science self-efficacy level increased gradually from Level 1 to Level 4, which also provided more evidence for the structural validity of the instrument.

To further verify the construct validity of the instrument, evidence of triangulation through cross-validation was provided to supplement the interview evidence from the qualitative research. The interview results showed that 9 of the 12 students were competent in the corresponding science tasks and expressed incompetence for higher-level science tasks. However, 3 students felt they were not competent in the science tasks. Considering that tasks at the same level have varying difficulty, 2 additional science tasks at this level were added, and ultimately, the students expressed their perceived competence. Therefore, the results of the interviews as a whole provided evidence of structural validity.

Predictive validity

Pearson correlation coefficients between the science self-efficacy score and the standardized science academic test score were calculated. The results showed that there was a moderate positive correlation (0.36–0.75) between science self-efficacy and science academic achievements ($r = 0.58$). A simple linear regression found that science self-efficacy was a significant predictor of science academic achievements, accounting for 33.2% of the variance ($r = 0.577$, $R^2 = 0.332$, $p < 0.001$). This result is consistent with other researchers' findings (Honicke & Broadbent, 2016; Tuan et al., 2005). Thus, the results provided evidence of the predictive validity of the instrument.

The predictive validity of science self-efficacy at different levels was also examined separately. As shown in Table 7, the results indicate that as the science self-efficacy level improved, the predictive effect of science academic achievement was stronger, and higher science self-efficacy was correlated with improved science academic performance (Mataka & Kowalske, 2015; Tezer & Aşıksoy, 2015). However, Table 7 shows that the predictive effects of the different levels of science self-efficacy on science academic achievement were different. Science self-efficacy at Level 3 had the strongest predictive effect on science academic achievement ($r = 0.796^{***}$, $R^2 = 0.633$). Science self-efficacies below Level 1 ($r = 0.333$, $R^2 = 0.111$) and at Level 1 ($r = 0.254^{***}$, $R^2 = 0.064$) had a weak predictive effect on science academic achievement (Pajares, 1996).

Part 2: cross-sectional analysis

The second question of this study was how to use high school students' science self-efficacy level to explore and explain changes across grades and gender differences in science self-efficacy; this question was explored through a cross-sectional study. The ICC of the school level is 0.015 ($p = 0.614$), that of the grade level is 0.024 ($p = 0.683$) and that of the class level is 0.090 ($p = 0.311$), all of which are less than 0.100 ($p > 0.05$), indicating that the differences between the schools, grades and class levels are not significant and that the potential dependency issue can be ignored (Goldstein, 2011; Raudenbush & Bryk, 2002).

The normal distribution of the dependent variable residuals and equality of variance across the groups almost satisfied the basic assumptions of the ANOVA (skewness < 1; kurtosis < 1; Shapiro–Wilk test $p > 0.05$; Levene's test $p > 0.05$). The two-way ANOVA results are shown in Table 8. Science self-efficacy was significantly dependent on gender [F (1, 1558) = 4.68, $p < 0.05$] and grade [F (2, 1558) = 2.45, $p < 0.05$]. Moreover, there was no significant interaction between gender and grade in influencing science self-efficacy [F (2, 1558) = 1.901, $p > 0.05$]. The results of a least squares difference (LCD) post hoc comparison indicated significant differences between grades 10 and 11 ($p = 0.004 < 0.05$),

Table 7 Correlation and simple linear regression analyses between science self-efficacy and science academic achievement

	F	*p*	Pearson's *r*	R^2	Adjusted R^2	*N*
Total	$F(1,1562) = 777.611$	0.000	0.577***	0.332	0.332	1564
Below Level 1	$F(1,21) = 2.613$	0.121	0.333	0.111	0.068	23
Level 1	$F(1,634) = 43.543$	0.000	0.254***	0.064	0.063	636
Level 2	$F(1,413) = 161.639$	0.000	0.530***	0.281	0.280	415
Level 3	$F(1,354) = 611.438$	0.000	0.796***	0.633	0.632	356
Level 4	$F(1,132) = 25.675$	0.000	0.445***	0.198	0.192	134

***$p < 0.001$

Table 8 Two-way ANOVA results for gender- and grade-based science self-efficacy variations

Source of variation	ST	df	MS	*F*	*p*
Gender	4.697	1	4.697	10.532	0.001
Grade	4.845	2	2.422	5.432	0.004
Grade × gender	1.695	2	0.848	1.901	0.150
Error	694.751	1558	0.446		
Total	843.202	1564			

ST squares total, *df* degrees of freedom, *MS* mean of squares

as well as between grades 11 and 12 ($p = 0.019 < 0.05$), but no significant difference between grades 10 and 12 ($p = 0.418 > 0.05$).

A Chi-square test was used to explore whether the distribution of science self-efficacy levels differed among students by gender and grade. The Chi-square test, shown in Table 9, revealed significant differences in gender (Chi-square = 43.673, $p = 0.000 < 0.001$) and grade (Chi-square = 19.078, $p = 0.014 < 0.05$) among the four science self-efficacy levels. Overall, the development of students' science self-efficacy from grade 10 to grade 12 was a complex process in which science self-efficacy first decreased and then increased. From Fig. 4, it can be seen that from 10 to 11th grade, there was a more obvious increasing trend in the proportion of students with lower levels of science self-efficacy, such as Level 1 and below Level 1, while from 11 to 12th grade, the proportion of students with lower levels showed a more obvious decreasing trend. Even though the proportion decreased, it was higher than that in grade 10. Level 2 showed a trend of decreasing first in grades 10–11 and increasing again in grades 11–12, but after the increase, the proportion of those in Level 2 in grade 12 was not as high as that in grade 10. However, the number of students at higher levels, such as Levels 3 and 4, did not change appreciably from 10 to 12th grade. The proportion of students at Level 1 was the highest, approximately 50%, which was the direct reason for the complex change process in science self-efficacy among students in grades 10–12.

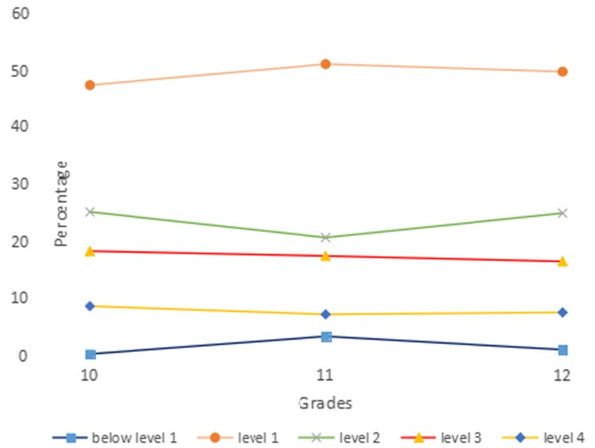


Fig. 4 Cross-sectional differences in science self-efficacy levels in grades 10–12

The distribution of gender differences in levels of science self-efficacy by mean score across grades is shown in Fig. 5. In general, boys had higher science self-efficacy than girls at all stages in grades 10–12. However, for different levels, there were large differences in science self-efficacy between boys and girls. In the lower levels, such as below and at Level 1, the science self-efficacy of most girls was greater than that of boys (almost no girls were below Level 1). At the highest level, Level 4, the science self-efficacy of girls was also greater than that of boys. However, in the middle levels, i.e., Levels 2 and 3, the science self-efficacy of boys was generally greater than that of girls.

Part 3: longitudinal analysis

The third problem addressed in this study was tracking how the science self-efficacy of the same students changed from grades 10 to 11 in China, which was investigated through a longitudinal study.

Students were tested in a 1-year follow-up from 10 to 11th grade. To develop an in-depth and comprehensive understanding of the problem, both qualitative and

Table 9 Distribution of science self-efficacy levels by gender and grade and Chi-square test

		Below Level 1 (%)	Level 1 (%)	Level 2 (%)	Level 3 (%)	Level 4 (%)	χ^2
Gender	Male	22 (3.15)	261 (37.40)	170 (24.35)	164 (23.50)	81 (11.60)	43.67***
	Female	1 (0.12)	375 (43.30)	245 (28.29)	192 (22.17)	53 (6.12)	
Grade	10	1 (0.29)	160 (47.48)	85 (25.22)	62 (18.4)	29 (8.61)	19.08*
	11	22 (3.36)	335 (51.14)	136 (20.76)	115 (17.56)	47 (7.18)	
	12	6 (1.05)	285 (49.82)	143 (25.00)	95 (16.61)	43 (7.52)	

*$p < 0.05$; ***$p < 0.001$

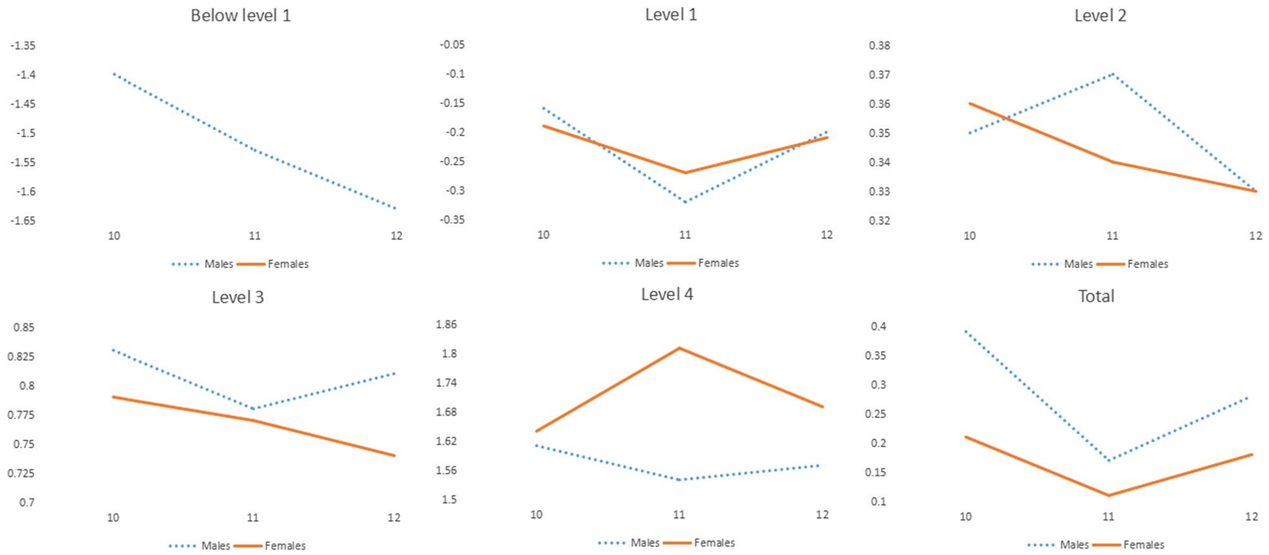


Fig. 5 Cross-sectional differences in science self-efficacy at each level by gender

quantitative data were collected (Creswell & Plano Clark, 2011). The quantitative data showed that the average science self-efficacy score of grade 10 students was 0.37 and that the average science self-efficacy score of the students followed up in grade 11 was 1.03. Moreover, the matched-samples *t* test found that there were significant differences in science self-efficacy between grades 10 and 11 ($t=7.79$, $p=0.000$). In addition, the effect size, namely, Cohen's *d* value, was 0.75, indicating a large effect (Valentine & Cooper, 2003).

From grades 10–11, the science self-efficacy of 39 students (16.74%) decreased, and that of 194 students (83.26%) increased. The changes in the proportions of students with science self-efficacy at each level are shown in Fig. 6. The proportion of students with a high level of science self-efficacy, such as Level 4, increased, while the proportion of students with a low level of science self-efficacy, such as Level 1, increased. The Chi-square analysis of the longitudinal study data is presented in Table 10 (only 3 students below Level 1 were in grade 10, and none were in grade 11, so they were not included in the statistics). The χ^2 value was 100.929, $p=0.000$, and the Chi-square results indicated a significant difference between 10 and 11th grade students in the change in the level of science self-efficacy.

To obtain detailed information that could not be captured by quantitative research, qualitative research was also conducted (Mataka & Kowalske, 2015). The qualitative data were derived from semistructured interviews, which included 2 main questions. The semistructured interviews allowed the interviewer to ask questions other than those on the interview outline and clarify the answers and to understand the real levels of and changes in students' science self-efficacy. The interview recordings of students were transcribed, and all interview records were provided to the participants to ensure that the interview information accurately reflected their real ideas. The information of the interviewees is shown in

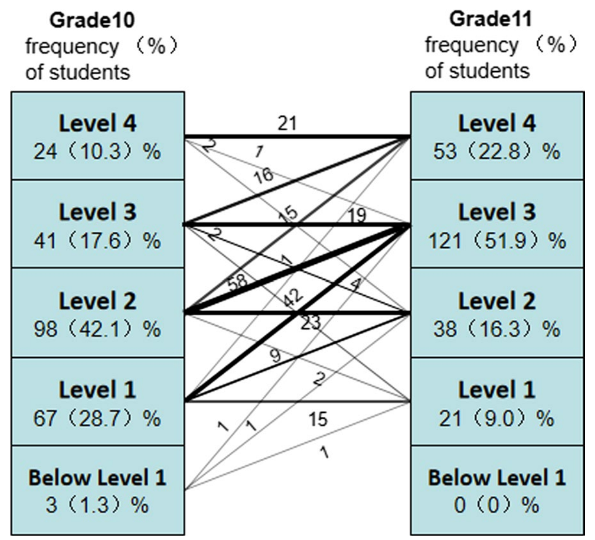


Fig. 6 Changes in the proportions of the students at each level of science self-efficacy (the thickness of the line indicates the number of students who changed)

Table 10 Chi-square contingency table of grade and science self-efficacy level

		Level 1 (%)	Level 2 (%)	Level 3 (%)	Level 4 (%)	χ^2	*p*
Grade	10	67 (28.75)	98 (42.06)	41 (17.59)	24 (10.30)	100.929	0.000
	11	21 (9.01)	38 (16.31)	121 (51.93)	53 (22.75)		

Table 11 Information about the level changes of the interviewees

Student	Gender	Measure in 10th	Level in 10th	Measure in 11th	Level in 11th	Changes
WGY12	Male	− 1.56	Below 1	0.77	3	Growth
HMZX01	Male	0.04	1	0.98	3	Growth
SSY75	Female	− 0.29	1	− 0.81	1	Reduction
LVXX18	Female	0.33	2	1.39	4	Growth
SSY70	Male	0.27	2	− 1.05	1	Reduction
JNZX02	Female	0.71	3	1.55	4	Growth
WGY04	Male	0.71	3	− 0.2	1	Reduction
JNZX36	Male	1.48	4	2.95	4	Growth
SSY35	Male	1.16	4	0.77	3	Reduction

Table 11. The specific interview questions from the semi-structured interviews were as follows: (1) Do you think you are more or less confident about your competence in science tasks after a 1-year science class? (2) What kind of science tasks in the class have increased or reduced your perceived competence in science tasks, and can you give an example?

Regarding the first question, some students indicated that after 1 year of high school science learning, science self-efficacy improved to a certain extent. Student JNZX36 mentioned, "I think I am more confident. I have always been very interested and confident in science. High school science courses have not baffled me, so I believe I can learn science well." Student LVXX18 stated, "I feel more confident in learning science. Although my science achievements have not improved too much, I find myself doing very well in science classes, and teachers often praise my ideas and design schemes." However, some students said that their science self-efficacy decreased. Student WGY04 reported, "I feel less confident in science tasks. I have not done these science inquiry experiments in person before. In junior high school, I watched all kinds of science experiments that my teacher played on videos. I thought it was very simple. I could do the same. However, in high school, when I personally designed and implemented the experiment, I found it too difficult to conduct the science inquiry experiment."

Regarding the second question, most students considered changes in science self-efficacy to be related to the experience of the science inquiry task and the interpretation of their performance in completing the science task. Student WGY12 stated, "Doing science experiments by myself successfully makes me very excited and allows me to form a deeper impression of science principles, which made me have more confidence in science." Student JNZX02 stated, "Although I failed in the science experiment, I know the reason for my failure. It's my carelessness. This is the biggest lesson of my class. I believe I will succeed next time." In contrast, student SSY75 stated, "I'm not confident in science all the time. I can only remember some basic science knowledge. Although I can get good marks on the science exam, if I am asked to solve a science problem in science class or daily life, I cannot do it."

Discussion

High school is a key period for the development of science self-efficacy. The accurate characterization of the level of students' science self-efficacy is a precondition for cultivating this perception (Aydın & Uzuntiryaki, 2009). Therefore, this study focused on developing a scale based on the Rasch model that could measure high school students' science self-efficacy focusing on perceptions of their ability to complete science tasks. The instrument consists of 24 items designed to divide students' science self-efficacy into 4 levels: identifying and remembering, explaining and applying, inquiring and argument, and innovative design.

The Rasch model provided an effective analytical theory and framework for the development and validation of the SSES (Boone et al., 2011). The validity of the instrument was verified from the aspects of face, content, construct, and predictive validity. To verify face validity, pairwise comparisons were first conducted to allow students to perceive and compare the difficulty of science tasks and to ensure the rationality of subsequent construct validity and predictive validity results (DeVellis, 2017; Liu, 2010). Second, with the help of expert group interviews, the content validity of the SSES was determined. Third, 373 high school students and 1564 students were tested in a pilot study and field test, respectively, and the construct validity was further verified by means of unidimensionality, local independence, and the Wright map. Fourth, the correlation analysis and simple linear regression with science academic achievement verified the predictive validity of the SSES. This verification of multiple aspects of validity can be regarded as interdependent and complementary from many aspects. One type of validity cannot replace the effectiveness of another, thus necessitating a comprehensive verification of instrument validity (Messick, 1989) to ensure the quality of the scale to the greatest extent.

A two-way ANOVA in the cross-sectional study found that the change in science self-efficacy differed significantly across grades and genders, but there was no interaction effect between grades and genders. Changes in science self-efficacy during grades 10–12 are the result of complex processes found in cross-sectional research. The change in high school students' science self-efficacy does not show a trend of constantly decreasing, as previous research has found (e.g., Eccles et al., 1997). Instead, it shows a process of decreasing and then increasing, with the final result still being lower than the initial results. On the one hand, in Chinese high schools, grade 12 students no longer learn new science knowledge but, instead, undergo a review of the science knowledge gained in grades 10 and 11. There is a significant positive correlation between proficiency in tasks or knowledge and self-efficacy (McCoy, 2010). Therefore, the improvement of students' proficiency in science knowledge may be an important reason for the development of science self-efficacy from grades 11 to 12. On the other hand, the proportion of students at lower levels, such as at Level 1 and below Level 1, changes greatly with the increase in grade, while the proportion of students at higher levels, such as Levels 3 and 4, remains almost unchanged. However, the proportion of students in Level 1 is almost 50%, which also affects the grade change trend in science self-efficacy.

In addition, from the perspective of the differences in the level of science self-efficacy between boys and girls according to Chi-square analysis, this study contributes to the debate on the gender differences in science self-efficacy found in the literature. Some researchers have found that girls have higher science self-efficacy than boys (Britner & Pajares, 2001; Pajares et al., 2000), and others have found that girls have lower science self-efficacy than boys (Chan, 2022; Weisgram & Bigler, 2006). Furthermore, some researchers have reported that there is no difference in science self-efficacy between boys and girls (Lips, 1992; Rowe, 1988; Sezgintürk & Sungur, 2020). Cross-sectional research on male and female students at various levels found that although male students had higher science self-efficacy than female students as a whole, at lower levels of below Level 1 and Level 1 and at the highest level, i.e., Level 4, female students had higher science self-efficacy than male students. For the middle levels, including Levels 2 and 3, male students had higher science self-efficacy than female students.

The reason why previous studies have drawn different conclusions may be related to sampling. The final conclusions were controversial due to the different proportions of boys and girls at each level. For example, if most of the selected students were at Levels 2 or 3, this may have led to the conclusion that the science self-efficacy of boys is higher than that of girls. In contrast, if most of the selected students were at Levels 1 or 4, it may have been concluded that the science self-efficacy of girls is higher than that of boys. The implications of the above conclusions for teaching are that the SSES is designed not only to diagnose students at low levels of intervention but also to focus particularly on male students at low and high levels and female students at intermediate levels to promote the overall development of student science self-efficacy.

A longitudinal follow-up study of 233 students from grades 10 to 11 showed that students' science self-efficacy greatly improved. The results of the matched-samples *t* test indicated that there was a significant difference in students' science self-efficacy between grades 10 and 11 with a large effect size. The results of the Chi-square test indicated significant differences between 10 and 11th grade students in the levels of science self-efficacy. It can be intuitively seen from Fig. 6 that the proportion of students in the lower levels, such as Levels 1 and 2, showed an upward trend, while the proportion of students in Level 4 also rose steadily.

However, the results of the longitudinal study contradict the findings of the cross-sectional study that the science self-efficacy in grades 10–11 decreased significantly. An important reason for the contradiction may be the effect of the science curriculum; the largest difference in the selected sample in the two-part study is students' experience with different science curricula. The science curriculum used in the cross-sectional study has been continuously implemented in China since 2000, and the core competency-oriented science curriculum used in the longitudinal study was officially implemented in China in September 2019. The core competency-oriented science curriculum emphasizes allowing students to experience science inquiry, creative thinking and critical thinking and strengthens their attention to science experiments (e.g., it adds 9 compulsory experiments in chemistry and 12 compulsory experiments in physics) (Yao & Guo, 2018), which are closely related to the Level 3 and 4 science tasks, providing students with successful experiences or the opportunity to observe others' successful experiences. These experiences are important sources of self-efficacy (Bandura, 1997). In addition, the important role of science inquiry and critical thinking in promoting the development of science self-efficacy has provided abundant empirical evidence in previous studies (e.g., Dehghani et al., 2011; Jansen et al., 2015).

The qualitative interview data indicated that, in comparison to providing students with others' successful experiences through watching a teacher's demonstration or instructional video in a traditional science curriculum, students perceived that experiencing science tasks in person was more beneficial in enhancing their science self-efficacy, which is consistent with the findings of other researchers (Bandura, 1997; Webb-Williams, 2018). Moreover, changes in science self-efficacy do not exclusively depend on completing the science task experience and may also be related to how students interpret their experiences (Bandura, 1997, 2012). For example, even though student JNZX02 failed a complex science task, she increased her science self-efficacy by positively interpreting this science task experience. The present study's findings must be regarded with caution, however. No randomized control group was used, nor was a comparison group used for either the surveys or interviews.

Study limitations and future research

The samples studied were Chinese high school students; future research needs to consider a wider sample of students to better promote the SSES. Influenced by Chinese traditional culture and Confucianism, Chinese students are quite different from students from the United States, Europe and other Western countries (Morony et al., 2013). Future research should also expand the sample range and increase the number of represented countries and ethnicities or nationalities to ensure the effectiveness of the SSES. The instrument developed in this study is mainly intended for high school students, but the difficulty division of science tasks is not unique to high school, so it should also be extended to the junior high school and university contexts. This is also an important direction of future research to further verify the external validity of the instrument (Aydın & Uzuntiryaki, 2009). The longitudinal study reached an interesting conclusion, but multiple factors appeared to influence students' self-efficacy, such as internal personal factors and environmental events (Bandura, 1986, 1997). Therefore, future research will continue to focus on the factors influencing science self-efficacy, such as designing experiments or quasi-experiments to explore the effects of the core competency-oriented science curriculum on science self-efficacy.

Conclusions

This study developed an instrument to measure high school students' science self-efficacy focusing on perceived competence dimension, and verified the validity of the instrument from the aspects of face validity, content validity, construct validity, and predictive validity using mixed methods, providing rich and rigorous evidence through triangulation by cross-validation. In addition, a cross-sectional analysis was conducted to explore the trends in science self-efficacy across increasing grades as well as the differences by gender, and an attempt was made to provide a detailed analysis and explanation from the perspective of the science self-efficacy level. Finally, it was found that students' science self-efficacy significantly improved in the longitudinal study, which was explained by combining self-efficacy theory and the core competency-oriented science curriculum in China. In conclusion, this study provided an instrument with high reliability and validity for the measurement of high school students' science self-efficacy as well as some empirical evidence-based suggestions for the development of science self-efficacy.

Appendix

See Tables 12, 13, 14, 15.

Table 12 Levels of science tasks, relevant behavioral performance codes, and source examples

Levels of science tasks	Behavioral performance codes	Examples of sources

Identifying and remembering	Identifying information;
	Recognizing a problem;
	Observing a phenomenon;
	Remembering a fact;
	Recognizing a relationship;
	Recalling a fact;
	Describing information;
	Comparing phenomena;
	Finding a law;
	Classifying information
		Interview: "Identification of data and information in graphs or tables";
		PISA 2006: Recognizing issues that are possible to investigate scientifically;
		Interview: "Carefully observe color changes in a chemical experiment";
		Interview: "Remember the formulas in physics and chemical equations";
		Interview: "Recognizing that iron and dilute sulfuric acid can react chemically";
		NAEP: NAEP 2009-12S10#2;
		PISA 2003: Describing scientific phenomena
		NAEP 2011-8S11 #3;
		NAEP 2009-12S9 #14;
		Interview: "Classification of chemicals into compounds, single substances and other classes";

Explaining and applying	Explaining phenomena;
	Calculating variables;
	Explaining a purpose
	Describing examples;
	Interpreting a model;
	Predicting phenomena;
	Predicting knowledge;
	Building a model
	Explaining a relationship;
	Making judgments;
		NAEP 2009-12S10 #1
		NAEP 2009-12S10 #3
		Interview: "Explaining the purpose of controlling variables in photosynthesis experiments";
		Interview: "Giving an example of what a particle is";
		Interview: "Explaining the working principle model of a primary battery"
		NAEP: "Predicting the results of the observation"
		NAEP 2009-12S9 #16
		Interview: "Explaining phenomena by building a particle model, charge model, atomic structure model, etc."
		Interview: "Explaining the connection between atomic structure and properties of matter";
		Interview: "Judging the reducibility of zinc, iron and copper"

Inquiring and argument	Evidence-based argument;
	Rational evaluation;
	Systemic inquiring;
	Selecting instruments;
	Proposing programs
		TIMSS: "Proposing arguments based on the evidence";
		Interview: "Evaluating and demonstrating the rationality of different design schemes or assumptions";
		NAEP 2009-12S9 #7;
		NAEP: "Using appropriate tools and techniques to conduct scientific inquiry"
		PISA: "Proposing a feasible research program for a given question"

Innovative design	Weighing choices;
	Improving defects;
	Improving living;
	Designing and completing production
		NAEP: Identifying scientific trade-offs in design decisions and choosing among alternative solutions
		NAEP: Proposing or critiquing solutions to problems given criteria and scientific constraints
		TIMSS: "Proposing research questions based on observation"
		Interview: "Designing and making a simple alcohol detector based on chemical principles and sensors"

Table 13 Initial SSES

Directions:

We are conducting investigative research on the learning situation of science disciplines (physics, chemistry, biology) with the hope of improving teaching. The results of the questionnaire are for research purposes only and will not have any adverse impact on you. Please answer according to your actual situation and tick "√" the corresponding options. The following are the learning tasks encountered in science learning (including physics, chemistry and biology). Please evaluate them truthfully according to your real feelings in the learning process of high school.

THANKS FOR YOUR HELP!

School ________ Grade _____ Major_____ Gender ____ Student numbers ____

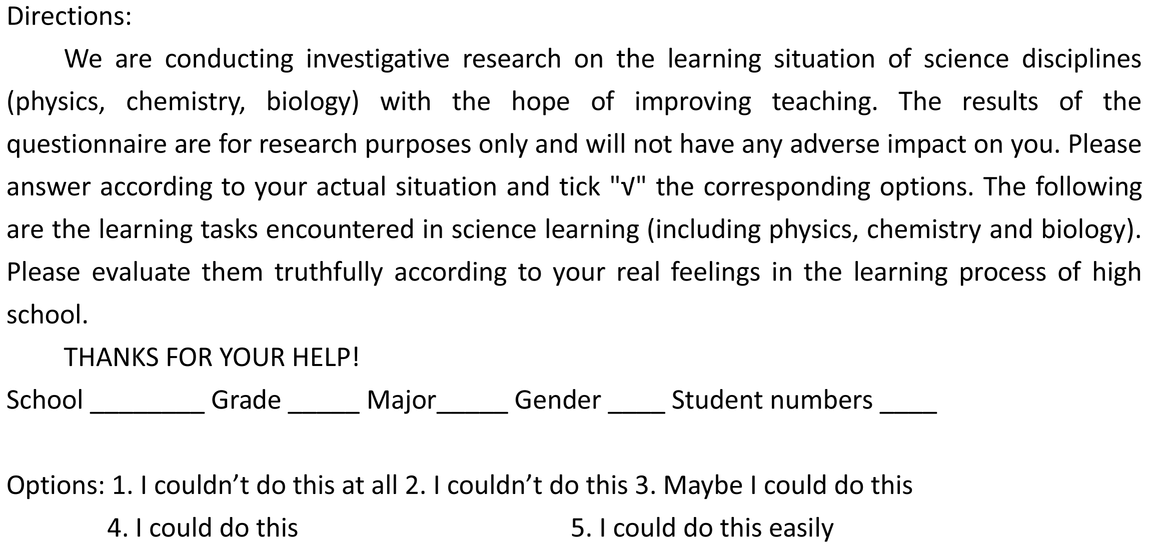
Options: 1. I couldn't do this at all 2. I couldn't do this 3. Maybe I could do this
4. I could do this 5. I could do this easily

Item no.	Item description	1	2	3	4	5
Q1	Recognizing information from text, images, graphs, and tables					
Q2	Recognizing variables that need to be studied in an experiment					
Q3	Observing and recording a phenomenon in an experiment and daily life					
Q4	Memorizing simple science concepts, principles, formulas, or equations					
Q5	Recognizing basic relationships between familiar things or concepts					
Q6	Recalling science concepts, principles, formulas, or equations when needed					
Q7	Accurately describing information observed or identified					
Q8	Identifying the similarities and differences between different science phenomena					
Q9	Finding science laws in experimental phenomena					
Q10	Classifying familiar things or science concepts					
Q11	Using scientific principles to explain daily life					
Q12	Calculating and comparing the unknown variables according to the given data, variables and other information by using certain science laws					
Q13	Interpreting the purpose of a specific operation in science experiments					
Q14	Describing an example to explain a science concept or principle					
Q15	Understanding and interpreting existing models					
Q16	Predicting an experimental phenomenon based on scientific principles					
Q17	Predicting new knowledge according to the relationship between science concepts					
Q18	Constructing models to explain science phenomena					
Q19	Explaining the relationship between different concepts or things					
Q20	Using scientific principles to make simple judgments about properties of or changes in matter					
Q21	Evaluating and demonstrating the reasonableness of different design options or assumptions					
Q22	Extracting evidence from the facts and proving or disproving a scientific hypothesis based on the evidence					
Q23	Selecting and integrating appropriate tools and instruments for scientific inquiry according to the purpose of the experiment					
Q24	Carrying out a complete and systematic inquiry to solve science problems					
Q25	Proposing one or more sets of research programs for a given science question					
Q26	Designing a science and technology program through weighing multifaceted and multiangled pros and cons					
Q27	Identifying defects in science equipment or daily objects and using practical design and transformation to improve them					
Q28	Identifying problems in everyday life and using science to design solutions					
Q29	Designing and completing a practical technology product independently based on scientific principles					

Table 14 Expert review outline

Part	Interview content
Part 1	Discuss whether the item corresponds to the science self-efficacy and corresponding task level shown in Fig. 1 If not, please give your comments and suggestions
Part 2	Discuss whether the items in each level are sufficient to represent the measure of students' science self-efficacy What other items should be included or excluded in the current version? Why or why not? Please provide comments and suggestions
Part 3	Discuss whether the item expression, wording, and language issues are appropriate 1. Were the item statements clearly expressed? Please point out the items and provide detailed comments 2. Was there any wording or language issues that might be confusing to you? Please point out these items and provide detailed comments 3. Are there any suggestions for the improvement of those item statements? Please point out these items and provide detailed comments
Part 4	Provide an overall evaluation and suggestions for further improvement of the instrument

Table 15 Final SSES



Directions:

We are conducting investigative research on the learning situation of science disciplines (physics, chemistry, biology) with the hope of improving teaching. The results of the questionnaire are for research purposes only and will not have any adverse impact on you. Please answer according to your actual situation and tick "√" the corresponding options. The following are the learning tasks encountered in science learning (including physics, chemistry and biology). Please evaluate them truthfully according to your real feelings in the learning process of high school.

THANKS FOR YOUR HELP!

School ________ Grade _____ Major_____ Gender ____ Student numbers ____

Options: 1. I couldn't do this at all 2. I couldn't do this 3. Maybe I could do this
4. I could do this 5. I could do this easily

Item no.	Item description	1	2	3	4	5
Q1	Recognizing information from text, images, graphs, and tables					
Q2	Observing a phenomenon in an experiment and daily life					
Q3	Memorizing simple science concepts, principles, formulas, or equations					
Q4	Recognizing basic relationships between familiar things or concepts					
Q5	Accurately describing information observed or identified					
Q6	Identifying the similarities and differences between different science phenomena					
Q7	Classifying familiar things or science concepts					
Q8	Using scientific principles to explain common phenomena in an experiment or daily life					
Q9	Calculating and comparing unknown variables according to the given data, variables and other information by using certain science laws					
Q10	Interpreting the purpose of a specific operation in science experiments					
Q11	Describing an example to explain a science concept or principle					
Q12	Understanding and interpreting existing models					
Q13	Predicting an experimental phenomenon based on scientific principles					
Q14	Predicting new knowledge according to the relationship between science concepts					
Q15	Using scientific principles to make simple judgments about properties of or changes in matter					
Q16	Evaluating and demonstrating the reasonableness of different design options or assumptions					
Q17	Extracting evidence from the facts and proving or disproving a scientific hypothesis based on the evidence					
Q18	Selecting and integrating appropriate tools and instruments for scientific inquiry according to the purpose of the experiment					
Q19	Carrying out a complete and systematic inquiry to solve science problems					
Q20	Proposing one or more sets of research programs for a given science question					
Q21	Designing a science and technology program through weighing multifaceted and multiangled pros and cons					
Q22	Identifying defects in science equipment or daily objects and using practical design and transformation to improve them					
Q23	Identifying problems in everyday life and using science to design solutions to improve the living environment					
Q24	Designing and completing a practical technology product independently based on scientific principles					

Abbreviations

STEM: Science, technology, engineering, and mathematics; OECD: Organisation for Economic Co-operation and Development; SSES: Science self-efficacy scale; ANOVA: Analysis of variance; ICC: Intra-class correlation; RMSEA: Root mean square error of approximation; SRMR: Standardized root mean residual; CFI: Comparative fit index; GFI: Goodness-of-fit index.

Acknowledgements

The authors thank all the schools who carried out the tests locally, the teachers and other experts who contributed to developing the tests, and the staff of the Education and Youth Authority who provided technical and other types of support in preparing and conducting the tests.

Author contributions

XH and YJ developed the survey instrument, conceptualized the study, and designed the methodology. XH performed the literature review, administered the survey, conducted the data analysis and was the primary manuscript author. HB advised on the strategy for the literature review. All the authors contributed to the manuscript writing and have read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Competing interests

The authors declare no competing interests.

Received: 6 January 2022 Accepted: 5 July 2022
Published online: 16 July 2022

References

Ainscough, L., Foulis, E., Colthorpe, K., Zimbardi, K., Robertson-Dean, M., Chunduri, P., & Lluka, L. (2016). Changes in biology self-efficacy during a first-year university course. *CBE Life Sciences Education, 15*(2), 1–12. https://doi.org/10.1187/cbe.15-04-0092

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., & Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (abridged). Longman.

Andrich, D. (1988). *Rasch models for measurement*. Sage.

Aydın, Y. Ç., & Uzuntiryaki, E. (2009). Development and psychometric evaluation of the high school chemistry self-efficacy scale. *Educational and Psychological Measurement, 69*(5), 868–880. https://doi.org/10.1177/0013164409332213

Bakeman, R., & Gottman, J. M. (1986). *Observing behavior: An introduction to sequential analysis*. Cambridge University.

Baldwin, J. A., Ebert-May, D., & Burns, D. J. (1999). The development of a college biology self-efficacy instrument for nonmajors. *Science Education, 83*(4), 397–408. https://doi.org/10.1002/(SICI)1098-237X(199907)83:4%3c397::AID-SCE1%3e3.0.CO;2-%23

Ballen, C. J., Wieman, C., Salehi, S., Searle, J. B., & Zamudio, K. R. (2017). Enhancing diversity in undergraduate science: Self-efficacy drives performance gains with active learning. *CBE Life Sciences Education, 16*(4), 1–6. https://doi.org/10.1187/cbe.16-12-0344

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*, 191–215. https://doi.org/10.1037/0033-295X.84.2.191

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist, 37*, 122–147. https://doi.org/10.1037/0003-066X.37.2.122

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory. Englewood Cliffs*. Prentice Hall.

Bandura, A. (1994). Self-Efficacy. In V. S. Ramachaudran (Ed.), *Encyclopedia of Human Behavior 4* (pp. 71–81). Academic Press.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. W. H. Freeman and Company.

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 307–337). Information Age.

Bandura, A. (2012). On the functional properties of perceived self-efficacy revisited. *Journal of Management, 38*(1), 9–44. https://doi.org/10.1177/0149206311410606

Bejar, I. I. (1983). *Achievement testing: Recent advances*. Sage.

Blotnicky, K. A., Franz-Odendaal, T., French, F., et al. (2018). A study of the correlation between STEM career knowledge, mathematics self-efficacy, career interests, and career activities on the likelihood of pursuing a STEM career among middle school students. *International Journal of STEM Education*. https://doi.org/10.1186/s40594-018-0118-3

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurements in the human sciences* (2nd ed.). Lawrence Erlbaum Associates Inc.

Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: how different are they really? *Educational Psychology Review, 15*, 1–40. https://doi.org/10.1023/A:10213.02408382

Boone, W. J., & Staver, J. R. (2020). *Advances in Rasch analyses in the human sciences*. Springer.

Boone, W. J., Townsend, J. S., & Staver, J. (2011). Using Rasch theory to guide the practice of survey development and survey data analysis in science education and to inform science reform efforts: An exemplar utilizing STEBI self-efficacy data. *Science Education, 95*(2), 258–280. https://doi.org/10.1002/sce.20413

Britner, S. L. (2008). Motivation in high school science students: A comparison of gender differences in life, physical, and earth science classes. *Journal of Research in Science Teaching, 45*(8), 955–970. https://doi.org/10.1002/tea.20249

Britner, S. L., & Pajares, F. (2001). Self-efficacy beliefs, motivation, race, and gender in middle school science. *Journal of Women and Minorities in Science and Engineering, 7*(4), 271–285. https://doi.org/10.1615/JWomenMinorScienEng.v7.i4.10

Britner, S. L., & Pajares, F. (2006). Sources of science self-efficacy beliefs of middle school students. *Journal of Research in Science Teaching, 43*(5), 485–499. https://doi.org/10.1002/tea.20131

Byars-Winston, A., Estrada, Y., Howard, C., Davis, D., & Zalapa, J. (2010). Influence of social cognitive and ethnic variables on academic goals of underrepresented students in science and engineering: A multiple-groups analysis. *Journal of Counseling Psychology, 57*, 205–218. https://doi.org/10.1037/a0018608

Bybee, R., McCrae, B., & Laurie, R. (2009). PISA 2006: An assessment of scientific literacy. *Journal of Research in Science Teaching, 46*(8), 865–883. https://doi.org/10.1002/tea.20333

Çalişkan, S., Selçuk, G. S., & Erol, M. (2007). Development of physics self-efficacy scale. *AIP Conference Proceedings., 899*(1), 483–484. https://doi.org/10.1063/1.2733247

Cassidy, S., & Eachus, P. (2002). Developing the computer user self-efficacy (CUSE) scale: Investigating the relationship between computer self-efficacy, gender and experience with computers. *Journal of Educational Computing Research, 26*(2), 133–153. https://doi.org/10.2190/JGJR-0KVL-HRF7-GCNV

Chan, R. C. (2022). A social cognitive perspective on gender disparities in self-efficacy, interest, and aspirations in science, technology, engineering, and mathematics (STEM): The influence of cultural and gender norms. *International Journal of STEM Education, 9*(1), 1–13. https://doi.org/10.1186/s40594-022-00352-0

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. Sage.

Cho, J. Y., & Lee, E-H. (2014). Reducing confusion about grounded theory and qualitative content analysis: Similarities and differences. *The Qualitative Report*, 19(64), 1–20. Retrieved from http://www.nova.edu/ssss/QR/QR19/cho64.pdf

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Sage.

Dalgety, J., & Coll, R. K. (2006). Exploring first-year science students' chemistry self-efficacy. *International Journal of Science and Mathematics Education, 4*(1), 97–116. https://doi.org/10.1007/s10763-005-1080-3

Dehghani, M., Pakmehr, H., & Malekzadeh, A. (2011). Relationship between students' critical thinking and self-efficacy beliefs in Ferdowsi University of Mashhad, Iran. *Procedia-Social and Behavioral Sciences, 15*, 2952–2955. https://doi.org/10.1016/j.sbspro.2011.04.221

DeVellis, R. F. (2017). *Scale development. Theory and applications* (4th ed.). Sage.

Duncan, P. W., Bode, R. K., Lai, S. M., Perera, S., & Glycine Antagonist in Neuroprotection Americas Investigators. (2003). Rasch analysis of a new stroke-specific outcome scale: the Stroke Impact Scale. *Archives of physical medicine and rehabilitation, 84*(7), 950–963. https://doi.org/10.1016/S0003-9993(03)00035-2.

Eccles, J. S., Midgley, C., Wigfield, A., Buchanan, C. M., Reuman, D., Flanagan, C., & Mac Iver, D. (1997). Development during adolescence: The impact of stage–environment fit on young adolescents' experiences in schools and in families (1993). In J. M. Notterman (Ed.), *The evolution of psychology: Fifty years of the American Psychologist* (pp. 475–501). American Psychological Association.

Erickson, F. (2012). Qualitative research methods for science education. *Second international handbook of science education* (pp. 1451–1469). Springer.

Finaulahi, K. P., Sumich, A., Heym, N., & Medvedev, O. N. (2021). Investigating psychometric properties of the self-compassion scale using Rasch methodology. *Mindfulness, 12*(3), 730–740. https://doi.org/10.1007/s12671-020-01539-8

Fiorella, L., Yoon, S. Y., Atit, K., et al. (2021). Validation of the Mathematics Motivation Questionnaire (MMQ) for secondary school students. *International Journal of STEM Education*. https://doi.org/10.1186/s40594-021-00307-x

Francis, J. J., Johnston, M., Robertson, C., Glidewell, L., Entwistle, V., Eccles, M. P., & Grimshaw, J. M. (2010). What is an adequate sample size? Operationalising data saturation for theory-based interview studies. *Psychology and Health, 25*(10), 1229–1245. https://doi.org/10.1080/08870440903194015

Gainor, K. A., & Lent, R. W. (1998). Social cognitive expectations and racial identity attitudes in predicting the math choice intentions of Black college students. *Journal of Counseling Psychology, 45*(4), 403–413. https://doi.org/10.1037/0022-0167.45.4.403

Gist, M. E., & Mitchell, T. R. (1992). Self-efficacy: A theoretical analysis of its determinants and malleability. *Academy of Management Review, 17*(2), 183–211. https://doi.org/10.5465/amr.1992.4279530

Glynn, S. M. (2012). International assessment: A Rasch model and teachers' evaluation of TIMSS science achievement items. *Journal of Research in Science Teaching, 49*(10), 1321–1344. https://doi.org/10.1002/tea.21059

Goldstein, H. (2011). *Multilevel statistical models* (4th ed.). John Wiley & Sons.

He, P., Zheng, C., & Li, T. (2021). Development and validation of an instrument for measuring Chinese chemistry teachers' perceived self-efficacy towards chemistry core competencies. *International Journal of Science and Mathematics Education*. https://doi.org/10.1007/s10763-021-10216-8

Heggestad, E. D., & Kanfer, R. (2005). The predictive validity of self-efficacy in training performance: Little more than past performance. *Journal of Experimental Psychology: Applied, 11*(2), 84–97. https://doi.org/10.1037/1076-898X.11.2.84

Hesse-Biber, S. N., & Johnson, R. B. (Eds.). (2015). *The Oxford handbook of multimethod and mixed methods research inquiry*. Oxford University Press.

Honey, M., Pearson, G., & Schweingruber, A. (2014). *STEM integration in K–12 education: Status, prospects, and an agenda for research*. National Academies Press.

Honicke, T., & Broadbent, J. (2016). The influence of academic self-efficacy on academic performance: A systematic review. *Educational Research Review, 17*, 63–84. https://doi.org/10.1016/j.edurev.2015.11.002

Huang, C. (2012). Gender differences in academic self-efficacy: A meta-analysis. *European Journal of Psychology of Education, 28*(1), 1–35. https://doi.org/10.1007/s10212-011-0097-y

Jansen, M., Scherer, R., & Schroeders, U. (2015). Students' self-concept and self-efficacy in the sciences: Differential relations to antecedents and educational outcomes. *Contemporary Educational Psychology, 41*, 13–24. https://doi.org/10.1016/j.cedpsych.2014.11.002

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14–26. https://doi.org/10.3102/0013189X033007014

Judge, T. A. (2009). Core self-evaluations and work success. *Current Directions in Psychological Science, 18*(1), 58–62. https://doi.org/10.1111/j.1467-8721.2009.01606.x

Kind, P. M. (2013). Conceptualizing the science curriculum: 40 years of developing assessment frameworks in three large-scale assessments. *Science Education, 97*(5), 671–694. https://doi.org/10.1002/sce.21070

Kıran, D., & Sungur, S. (2012). Middle school students' science self-efficacy and its sources: Examination of gender difference. *Journal of Science Education and Technology, 21*(5), 619–630. https://doi.org/10.1007/s10956-011-9351-y

Klassen, R. (2002). A question of calibration: A review of the self-efficacy beliefs of students with learning disabilities. *Learning Disability Quarterly, 25*(2), 88–102. https://doi.org/10.2307/1511276

Klein, J. (2014). Assessing university students' achievements by means of standard score (Z score) and its effect on the learning climate. *Studies in Educational Evaluation, 40*, 63–68. https://doi.org/10.1016/j.stueduc.2013.12.002

Lamb, R. L., Vallett, D., & Annetta, L. (2014). Development of a short-form measure of science and technology self-efficacy using Rasch analysis. *Journal of Science Education and Technology, 23*(5), 641–657. https://doi.org/10.1007/s10956-014-9491-y

Larose, S., Ratelle, C. F., Guay, F., Senécal, C., & Harvey, M. (2006). Trajectories of science self-efficacy beliefs during the college transition and academic and vocational adjustment in science and technology programs. *Educational Research and Evaluation, 12*(4), 373–393. https://doi.org/10.1080/13803610600765836

Linacre, J. M. (2011). *A User's Guide to Winsteps & Ministep: Rasch-Model Computer Programs. [version 3.72.0]*. Chicago: winsteps.com.

Lips, H. M. (1992). Gender-and science-related attitudes as predictors of college students' academic choices. *Journal of Vocational Behavior, 40*(1), 62–81. https://doi.org/10.1016/0001-8791(92)90047-4

Liu, X. (2010). *Using and developing measurement instruments in science education: A Rasch modeling approach*. Information Age Publishing, Inc.

Liu, X., & McKeough, A. (2005). Developmental growth in students' concept of energy: Analysis of selected items from the TIMSS database. *Journal of Research in Science Teaching, 42*(5), 493–517. https://doi.org/10.1002/tea.20060.

Livinti, R., Gunnesch-Luca, G., & Iliescu, D. (2021). Research self-efficacy: A meta-analysis. *Educational Psychologist, 56*(3), 215–242. https://doi.org/10.1080/00461520.2021.1886103.

Lu, S., & Bi, H. (2016). Development of a measurement instrument to assess students' electrolyte conceptual understanding. *Chemistry Education Research and Practice, 17*(4), 1030–1040. https://doi.org/10.1039/C6RP00137H

Lu, H., Jiang, Y., & Bi, H. (2020). Development of a measurement instrument to assess students' proficiency levels regarding galvanic cells. *Chemistry Education Research and Practice, 21*(2), 655–667. https://doi.org/10.1039/C9RP00230H

Luo, M., Sun, D., Zhu, L., & Yang, Y. (2021). Evaluating scientific reasoning ability: Student performance and the interaction effects between grade level, gender, and academic achievement level. *Thinking Skills and Creativity, 41*, 100899. https://doi.org/10.1016/j.tsc.2021.100899

Mangos, P. M., & Steele-Johnson, D. (2001). The role of subjective task complexity in goal orientation, self-efficacy, and performance relations. *Human Performance, 14*(2), 169–185. https://doi.org/10.1207/S15327043HUP1402_03

Marginson, S., Tytler, R., Freeman, B., & Roberts, K. (2013). *STEM: country comparisons: International comparisons of science, technology, engineering and mathematics (STEM) education. Final report*. Australian Council of Learned Academies, Melbourne, Vic. http://hdl.handle.net/10536/DRO/DU:30059041

Mataka, L. M., & Kowalske, M. G. (2015). The influence of PBL on students' self-efficacy beliefs in chemistry. *Chemistry Education Research and Practice, 16*(4), 929–938. https://doi.org/10.1039/C5RP00099H

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research, 49*, 305–328. https://doi.org/10.1080/00273171.2014.911075

McClelland, D. C. (1998). Identifying competencies with behavioral-event interviews. *Psychological Science, 9*(5), 331–339. https://doi.org/10.1111/1467-9280.00065

McCoy, C. (2010). Perceived self-efficacy and technology proficiency in undergraduate college students. *Computers & Education, 55*(4), 1614–1617. https://doi.org/10.1016/j.compedu.2010.07.003

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Morony, S., Kleitman, S., Lee, Y. P., & Stankov, L. (2013). Predicting achievement: Confidence vs self-efficacy, anxiety, and self-concept in Confucian and European countries. *International Journal of Educational Research, 58*, 79–96. https://doi.org/10.1016/j.ijer.2012.11.002

Neuendorf, K. A. (2017). *The content analysis guidebook*. Sage.

Oishi, S., Schimmack, U., Diener, E., & Suh, E. M. (1998). The measurement of values and individualism-collectivism. *Personality and Social Psychology Bulletin, 24*(11), 1177–1189. https://doi.org/10.1177/01461672982411005

Oon, P. T., & Fan, X. (2017). Rasch analysis for psychometric improvement of science attitude rating scales. *International Journal of Science Education, 39*(6), 683–700. https://doi.org/10.1080/09500693.2017.1299951

Organization for Economic Co-operation and Development (OECD). (2008). Encouraging student interest in science and technology studies. Global Science Forum. Retrieved from the internet December 9, 2019: https://www.oecd.org/publications/encouraging-student-interest-in-science-and-technology-studies-9789264040892-en.htm

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research, 66*, 543–578. https://doi.org/10.3102/00346543066004543

Pajares, F., Britner, S. L., & Valiante, G. (2000). Relation between achievement goals and self-beliefs of middle school students in writing and science. *Contemporary Educational Psychology, 25*(4), 406–422. https://doi.org/10.1006/ceps.1999.1027

Pajares, F., & Miller, M. D. (1994). The role of self-efficacy and self-concept beliefs in mathematical problem-solving: A path analysis. *Journal of Educational Psychology, 86*, 193–203. https://doi.org/10.1037/0022-0663.86.2.193

Pajares, F., & Schunk, D. H. (2001). Self-beliefs and school success: Self-efficacy, self-concept, and school achievement. In R. Riding & S. Rayner (Eds.), *International perspectives on individuals differences: Self perception* (pp. 239–266). Ablex Publishing.

Pedaste, M., Baucal, A., & Reisenbuk, E. (2021). Towards a science inquiry test in primary education: Development of items and scales. *International Journal of STEM Education*. https://doi.org/10.1186/s40594-021-00278-z

Planinic, M., Boone, W. J., Susac, A., & Ivanjek, L. (2019). Rasch analysis in physics education research: Why measurement matters. *Physical Review Physics Education Research, 15*(2), 020111. https://doi.org/10.1103/PhysRevPhysEducRes.15.020111

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage.

Rindermann, H. (2007). The g-factor of international cognitive ability comparisons: The homogeneity of results in PISA, TIMSS, PIRLS and IQ-tests across nations. *European Journal of Personality, 21*(5), 667–706. https://doi.org/10.1002/per.634

Robnett, R. D., Chemers, M. M., & Zurbriggen, E. L. (2015). Longitudinal associations among undergraduates' research experience, self-efficacy, and identity. *Journal of Research in Science Teaching, 52*(6), 847–867. https://doi.org/10.1002/tea.21221

Rowe, K. J. (1988). Single-sex and mixed-sex classes: The effects of class type on student achievement, confidence and participation in mathematics. *Australian Journal of Education, 32*(2), 180–202. https://doi.org/10.1177/000494418803200204

Scherbaum, C. A., Cohen-Charash, Y., & Kern, M. J. (2006). Measuring general self-efficacy: A comparison of three measures using item response theory. *Educational and Psychological Measurement, 66*(6), 1047–1063. https://doi.org/10.1177/0013164406288171

Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist, 26*, 207–231. https://doi.org/10.1080/00461520.1991.9653133

Sezgintürk, M., & Sungur, S. (2020). A multidimensional investigation of students' science self-efficacy: The role of gender. *İlkogretim Online-Elementary Education Online, 19*(1), 208–218. https://doi.org/10.17051/ilkonline.2020.653660

Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling: A Multidisciplinary Journal, 3*(1), 25-40. https://doi.org/10.1080/10705519609540027.

Tatar, N., Yıldız, E., Akpınar, E., & Ergin, Ö. (2009). A study on developing a self-efficacy scale towards science and technology. *Egitim Arastirmalari-Eurasian Journal of Educational Research, 36*, 263–280.

Tezer, M., & Aşıksoy, G. Y. (2015). Engineering students' self-efficacy related to physics learning. *Journal of Baltic Science Education, 14*(3), 311–326. https://doi.org/10.33225/jbse/15.14.311

Thomas, B., & Watters, J. (2015). Perspectives on Australian, Indian and Malaysian approaches to STEM education. *International Journal of Educational Development, 45*, 42–53. https://doi.org/10.1016/j.ijedudev.2015.08.002

Trochim, W., & Donnelly, J. (2006). *The research methods knowledge base* (3rd ed.). Atomic Dog Publishing.

Tuan, H., Chin, C., & Shieh, S. (2005). The development of a questionnaire to measure students' motivation towards science learning. *International Journal of Science Education, 27*, 639–654. https://doi.org/10.1080/0950069042000323737

Uzuntiryaki, E., & Aydın, Y. Ç. (2009). Development and validation of chemistry self-efficacy scale for college students. *Research in Science Education, 39*(4), 539–551. https://doi.org/10.1007/s11165-008-9093-x

Valentine, J. C., & Cooper, H. (2003). *Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes*. What Works Clearinghouse.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states* (National Institute for Science Education NISE Research Monograph No. 18). Madison: University of Wisconsin-Madison, National Institute for Science Education.

Webb-Williams, J. (2018). Science self-efficacy in the primary classroom: Using mixed methods to investigate sources of self-efficacy. *Research in Science Education, 48*(5), 939–961. https://doi.org/10.1007/s11165-016-9592-0

Weisgram, E. S., & Bigler, R. S. (2006). Girls and science careers: The role of altruistic values and attitudes about scientific tasks. *Journal of Applied Developmental Psychology, 27*(4), 326–348. https://doi.org/10.1016/j.appdev.2006.04.004

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. MESA Press.

Yao, J. X., & Guo, Y. Y. (2018). Core competences and scientific literacy: The recent reform of the school science curriculum in China. *International Journal of Science Education, 40*(15), 1913–1933. https://doi.org/10.1080/09500693.2018.1514544

Zi, Y. (2010). Objective measurement in psychological science: An overview of Rasch model. *Advances in Psychological Science, 18*(08), 1298–1305.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.