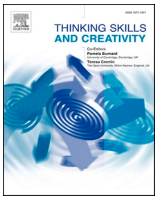
Contents lists available at ScienceDirect

Thinking Skills and Creativity

journal homepage: www.elsevier.com/locate/tsc



Development of the control of variables strategy in physics among secondary school students

De Van Vo[a,b,*], Benő Csapó[c,d,1,2], Samuel Greiff[e,1]

[a] An Giang University - VNU-HCM, 18 Ung Van Khiem St, Dong Xuyen Ward, Long Xuyen City, An Giang Province, Vietnam
[b] Doctoral School of Education, University of Szeged, 32-34. Petőfi S. sgt., Szeged H-6722, Hungary
[c] Institute of Education, University of Szeged, 32-34. Petőfi S. sgt., Szeged H-6722, Hungary
[d] MTA-SZTE Research Group on the Development of Competencies, Szeged, Hungary
[e] Department of Behavioural and Cognitive Sciences, University of Luxembourg, 4366 Esch-sur-Alzette, Luxembourg

ARTICLE INFO

Keywords:
Scientific reasoning
Control of variables strategy
Symmetric logistic function
Physics education
Rasch model

ABSTRACT

This study explores the developmental trend in scientific reasoning in the control of variables strategy (CVS) and how relevant factors contribute to explaining the individual abilities of secondary school students. A cross-sectional investigation involving 807 students from Grades 8 to 12 was conducted in eleven public schools in Vietnam. A 24-item test assessed CVS in basic physics (mechanics, thermodynamics, and electricity), emphasizing three CVS subskills (identifying controlled experiments, interpreting controlled experiment outcomes, and understanding the determinacy of confounded experiments). The results showed that the students' CVS capacity increased in a nonlinear pattern across grade levels with a model fitting in the symmetric logistic function, in which the most rapid growth was flagged in the second year of high school. Although there was no significant difference for gender in CVS either within each grade cohort or within the whole sample, the mean score favoured males on the item bundle of understanding the determinacy of confounded experiments. Furthermore, multi-model Bayesian inference suggested that grade level (or student age), prior content knowledge, and mother's education were better factors in predicting students' CVS capacities in this study. The implications for implementing the findings in educational practice are also discussed.

1. Introduction

Modern society is under pressure to process more information in a shorter amount of time. Instead of being taught a great deal of subject-specific information, students should thus be taught how to think critically and process information. Indeed, developing proficiency in scientific techniques has become a primary focus of science education programmes globally. As Lawson (2009) points out, the acquisition of both scientific reasoning and subject-specific knowledge in science contributes to the growth of scientific literacy. Specifically, the control of variables strategy (CVS) has been considered as a central scientific strategy in the context of scientific

[*] Corresponding author at: An Giang University-VNUHCM, 18 Ung Van Khiem St, Dong Xuyen Ward, Long Xuyen City, An Giang Province, Vietnam.

E-mail address: vvde@agu.edu.vn (D. Van Vo).

[1] Co-author.

[2] Benő Csapó deceased while the publication process of this paper was in its final stages. We mourn the loss of a wonderful person and a great researcher. This paper is dedicated to his memory.

https://doi.org/10.1016/j.tsc.2023.101371
Received 26 January 2022; Received in revised form 26 June 2023; Accepted 19 July 2023
Available online 20 July 2023

reasoning, which refers to conducting, predicting, and evaluating experimental systems (Chen & Klahr, 1999).

Scientific reasoning has been a major area of research in developmental psychology, with a significant focus on the development of CVS (Kuhn, 2007; Kuhn & Dean, 2005). Using cross-sectional data has the potential to offer deeper perspectives on the cognitive development of children in relation to CVS. This is crucial for education because CVS can be enhanced in school settings (Lorch et al., 2014, 2020; Schwichow et al., 2016), especially through inquiry-based instruction (Schlatter et al.,2020). However, it is most effective to stimulate thinking skills at the early stages of children's development, particularly during periods of rapid growth and at sensitive periods of development (Csapó, 1997). The developmental patterns of children's scientific reasoning, therefore, remain an interesting topic for psychologists and educators alike. A number of studies (Bullock & Ziegler, 1999; Ding, 2018; Han, 2013; Kwon & Lawson, 2000; Lazonder et al., 2021; Tairab, 2015) have investigated children's development in scientific reasoning, including CVS as an essential component of scientific reasoning. A study by Bullock and Ziegler (1999), a part of the Munich Longitudinal Study in scientific reasoning among students from three to twelve years old, examined the development of scientific reasoning and differences in individual development. The findings suggested that there was a significant increase in scientific reasoning abilities in the age range studied and that there were individual differences in this development. The results also showed that children's scientific reasoning abilities were influenced by both developmental and individual factors. A recent study by Schwichow et al. (2020), which involved secondary school students who were assessed on their use of CVS and their content knowledge in physics, pointed to a positive relationship between the use of CVS and physics content knowledge in that age group. The results demonstrated that promoting the use of CVS in physics can enhance students' content knowledge in that subject.

In Vietnam, experimental and inquiry activities are compulsory parts of the science curricula in secondary schools. Experimental activities account for a major proportion of programmes in the science subjects (MOET, 2009). Nevertheless, students' CVS scientific reasoning has rarely been involved on school or state examinations. Investigating CVS using cross-sectional methods may shed some light on the impact of current educational programmes on students' scientific reasoning abilities and increase recognition of the significance of promoting CVS in school practice and future curricular reforms. The current study investigates the developmental trend of students' control of variables strategy in physics (CVSP) by employing the symmetric logistic model with cross-sectional data. This study is expected to provide insight into developmental processes of CVS subskills, including identifying controlled experiments (ID), interpreting controlled experiment outcomes (IN), and understanding the determinacy of confounded experiments (UN) in children at the secondary education level. The study also considers relevant antecedents that contribute to individual students' CVSP ability. This is a preliminary step towards gaining a clearer understanding of science education in Vietnam within a global context. Specifically, the current study seeks to answer the following three research questions:

1 How can developmental curves in students' CVSP ability be characterized across grade cohorts?

   It is expected that students in the older cohort will exhibit significantly superior performance compared to those in the younger groups (Bullock & Ziegler, 1999; Han, 2013; Schwichow et al., 2020; Zimmerman, 2007).

2 Do males and females show significant disparities in their overall CVSP ability and in individual subskills?

   It is assumed that no significant difference will be found between males and females on the CVSP test (Mayer et al., 2014; Piraksa et al., 2014).

3 Which factors explain individual CVSP abilities among secondary school students?

   It is anticipated that individual CVSP capacity may be explained by the physics content test in the previous semester, by grade level (age range) (Schwichow et al., 2020), and by parents' education (Koerber et al., 2015).

2. Theoretical background

2.1. The control of variables strategy and its role in learning science

Content knowledge and scientific reasoning are two core pillars of scientific literacy (Lawson, 2009). Scientific reasoning involves reasoning and problem-solving skills with a series of cognitive and metacognitive processes to generate, test, and reassess hypotheses or theories (Zimmerman, 2007). CVS is the central component of scientific reasoning (Kuhn & Dean, 2005), which encompasses constructing arguments and drawing valid conclusions based on changes of individual variables in an experimental system (Boudreaux et al., 2008). As van der Graaf et al. (2015) discussed, CVS is defined in procedural and logical terms. It is a procedure for conducting experiments and distinguishing unconfounded experiments from confounded ones. Logically, CVS implies understanding the inherent indeterminacy of confounded experiments and making appropriate inferences from the outcomes of an experimental behaviour system. Several complex reasoning schemes and strategies are required to solve a CVS problem (Adey & Csapó, 2012). The rational principle of CVS in an experimental system requires varying one thing at a time. Recognizing variables that are testable and those that are constant in an experimental system plays a decisive role in handling a CVS task.

Furthermore, based on the definition of CVS by Chen and Klahr (1999), Schwichow et al. (2016) categorized CVS into four subskills: planning controlled experiments (PL), identifying controlled experiments (ID), interpreting controlled experiments (IN), and understanding the indeterminacy of confounded experiments (UN). A PL task evaluates a proposed experimental system with manageable and observable variables that students suggest based on given materials. An ID task typically starts with a stem, such as a brief narrative, which presents a hypothesis that needs to be tested. The task requires students to differentiate between variables in an experimental system that are testable and those that have causal effects in order to choose the appropriate ones. IN tasks are used to evaluate students' skills in deducing the results of a controlled experiment. They consist of a stem that depicts the proposed experimental set-up. Students must make appropriate inferences from the experimental set-up to arrive at valid conclusions. UN tasks are designed to assess students' deep understanding of the unreliability of experiments that are confounded. Their structure is similar to that of IN tasks; however, instead of drawing conclusions, students must determine if a given experiment is capable of yielding valid results (Schwichow et al., 2016; Van Vo & Csapó, 2021).

CVS is an acquisitive contribution to the development of scientific reasoning skills because it comprises inquiring about the components of experimental systems (Chen & Klahr, 1999) and emphasizes both engineering and scientific goals (Schauble et al., 1991). Students can avoid misconceptions in designing and interpreting experiments by learning from their mistakes when solving CVS problems (Siler & Klahr, 2012). In addition, Chen and Klahr (1999) found that students who are taught how to handle CVS tasks can apply their skills broadly to explore the relationship between different values of causal variables. Students' proficiency in CVS may indicate their prior subject-specific knowledge and aid in the development of additional content knowledge (Edelsbrunner et al., 2018, 2022; Stender et al., 2018). CVS can be enhanced in school settings through daily science classes with appropriate instruction (Lorch et al., 2014, 2020) and through school labs (Schlatter et al., 2020; van der Graaf et al., 2015; Wood et al., 2018), where students are engaged in inquiry activities under both implicit and explicit curricular conditions (Strand-Cary & Klahr, 2008; Vorholzer et al., 2020). Tytler and Peterson (2003) and Wood et al. (2018) recommend that schools should consider developing this skill for students at various educational levels. The main goals of STEM subjects, therefore, tend to be to acquire content knowledge and develop scientific reasoning. Kuhn and Dean (2005), however, have pointed out that too much importance should not be placed on control of variables, nor should other aspects of the process of scientific inquiry be ignored. They found that the initial question-formulating phase played a vital role in scientific inquiry.

2.2. Development of CVS across grade levels

Various empirical studies have shown that children's scientific reasoning capacity develops through the various educational levels. A study by van der Graaf et al. (2015) demonstrated that kindergarteners can use certain basic CVS strategies to explore the physical world around them. Children start developing some experimenting and conclusion-drawing skills at pre-school age (Koerber et al., 2015; Piekny et al., 2014). They can evaluate some relevant evidence to make some simple decisions at age four and develop a greater understanding of experimentation at the ages of five and six, but their grasp of experimentation remains at a low level at ages seven (Chen & Klahr 1999) and ten (Schauble, 1996).

Additionally, a longitudinal investigation by Bullock and Ziegler (1999) found that ID develops early in most children: between the third and fourth grades. By the third grade, children grasp some basic principles of experimentation, and they understand the need to control other variables by the fourth grade. The ability to plan experiments generally develops in most children around ages 10–12, while UN is likely the only skill that still requires development in many teenagers. A previous study by Han (2013), which involved the Lawson Classroom Test of Scientific Reasoning (LCTSR) to measure scientific reasoning, showed that the most rapid development of students' CVS capacity (ID and IN) occurred between the 10th and 11th grades. Furthermore, a cross-sectional investigation by Ding (2018) using LCTSR to assess students from elementary school to university demonstrated that students' scientific reasoning capacity increases grade by grade and apparently plateaus in the college years. The subskills, however, developed notably during the secondary education level but at different rates. The patterns of the cross-grade progression trends of the subskills are similarly low and relatively stable. Although CVS (i.e., ID and IN) was higher than the other scientific reasoning subskills during the early grade levels (4th–8th grades), the growth of the trend line started accelerating at a time point later than the others. Specifically, there was a noticeable increase in the 8th–9th grades for proportional reasoning and in the 9th–10th grades for probabilistic and correlational reasoning, but this was observed for CVS in the 10th–11th grades. A recent study by Schwichow et al. (2020) found that students developed from Grades 5 to 13. The students' scores on the PL, ID, and IN subtests grew through the lower secondary school years (the 5th–9th grades), while their performance on the UN scale rose between the 10th and 11th grades.

2.3. Gender difference in scientific reasoning

Previous studies have shown diverse findings on gender differences in scientific reasoning. Some studies in Germany (Koerber et al., 2015; Mayer et al., 2014), Thailand (Piraksa et al., 2014), and Finland (Thuneberg et al., 2015) have indicated that no significant difference exists between males and females, while other studies in China (Luo et al., 2021), the United Arab Emirates (Tairab, 2015), and Turkey (Tairab, 2015) have found that males performed better than females. Similarly, there have been divergent findings on CVS with regard to gender. Some research has demonstrated a significant difference between boys and girls, with boys being favoured by the results (e.g., Tairab, 2015; Tekkaya & Yenilmez, 2006; Valanides, 1997). However, other studies have shown that boys did not differ from girls on CVS tests (e.g., Mayer et al., 2014; Piraksa et al., 2014; Thuneberg et al., 2015). These inconsistent findings may be attributable to certain factors, such as different cultural contexts.

2.4. Factors predicting cvs

Both physical maturity and social experience (e.g., school and family environments) influence the development of children in thinking and reasoning capacities (Kwon & Lawson, 2000). Age (or grade level) is associated with scientific reasoning ability (Stevenson et al., 2013; Wagensveld et al., 2015) and is one of the main predictors of CVS (e.g., van der Graaf et al., 2015; Van Vo & Csapó, 2020, 2021a). The development of children's ability to experiment is closely linked to their understanding of knowledge, or epistemological understanding, which is in turn dependent on cognition as a foundation (Osterhaus et al., 2017). Existing findings (e.g., Han, 2013; Schwichow et al., 2020; Tairab, 2015) show that students' scientific reasoning develops across grade levels, but the growth rate depends on specific age ranges.

Children's CVS capacity can be promoted in school contexts (Lorch et al., 2014, 2020; M. Schwichow et al., 2016). Explicit instructions can effectively enhance students' CVS (Wagensveld et al., 2015), especially with the inquiry-based learning approach (Schlatter et al., 2020). CVS is also linked to children's achievement in learning science (Osterhaus et al., 2017; Song & Black, 1992), biology (Thompson et al., 2018), and physics (Coletta & Phillips, 2005; Schwichow et al., 2020; Van Vo & Csapó, 2021b). Conversely, learning school subjects (i.e., physics) meaningfully impacts on students' ability to solve CVS problems in secondary school (Bao et al., 2009). Prior content knowledge and CVS are closely related in secondary school contexts (e.g., Hejnová et al., 2018; Schwichow et al., 2020; Stender et al., 2018; Valanides, 1997).

CVS is closely tied to cognitive ability in general (Edelsbrunner et al., 2022). Studies by Koerber et al. (2015) and Molnár et al. (2013) showed that there is a strong link between students' scientific reasoning and general intelligence. The use of CVS is positively related to nonverbal reasoning ability (van der Graaf et al., 2015). Even in same-age children, scientific reasoning develops according to different patterns, largely independent of personal characteristics, but it is generally linked to other cognitive abilities (e.g., reading) (Koerber et al., 2015; Lazonder et al., 2021).

Spruijt et al. (2020) argue that educating parents may be a crucial asset in developing children's scientific and social reasoning, while Koerber et al. (2015) found that scientific thinking in children was significantly influenced by parental education. Also, school performance and parents' education are apparently related to students' scientific reasoning (Van Vo & Csapó, 2020, 2021a, 2023).

3. The study context

There are four levels in the Vietnamese national education system: early childhood education, general education (primary education, lower secondary education, and upper secondary education), vocational education, and higher education (Vietnam National Assembly, 2006). Children aged 11 begin four years (6th to 9th grades) in lower secondary education, while upper secondary education is for 15- to 18-year-old students. To continue to upper secondary education, 9th-grade students must achieve a sufficient score on a provincial selection examination. Likewise, in order to apply to university or college, all high school graduates must pass the National High School Graduation Examination after completing the 12th grade. The current general education curricula in the study were introduced nationwide during the 2002–03 school year, and they comprise the same objectives, contents, curriculum, and textbooks used in all public institutions across the country (UNESCO, 2011). Students learn science in different subjects from the 6th grade. Thinking and reasoning are incorporated into the core curricula, and experimental activities account for a major proportion of the science subject programmes. For example, the national education programme requires around 10% of the physics curriculum in high schools to include experimental lab activities (MOET, 2009).

Research on an effective developmental pattern is usually conducted using longitudinal studies (Becker et al., 2010; Ifenthaler & Seel, 2011; Lazonder et al., 2021; Strand-Cary & Klahr, 2008) and cross-sectional data (Csapó, 1997; Han, 2013; Koerber et al., 2015; Molnár et al., 2013; Van Vo & Csapó, 2020). A study spanning several years is quite challenging because it requires stable, long-term funding. However, cross-sectional data may acceptably estimate the real developmental pattern if environmental conditions change relatively slowly compared to the developmental process under observation. This cross-sectional investigation is expected to provide evidence to partially estimate how CVSP capacity develops and how well physics curricula enhance scientific reasoning abilities among secondary school students.

4. Methods

4.1. Participants

This study involved 807 students from the 8th to 12th grades in eleven public schools in Vietnam, with a mean age of 15.5 years (Table 1). Our goal was to ensure that each cohort correctly represented the grade level by involving two different schools (one from a city centre and another from the outskirts) with at least two classes per school. We used probability sampling based on clusters of 160

Table 1
Participants.

Grade	*n*	Male/Female Ratio (%)	Mean age (years)	Age range	No. of classes
8	178	43.3/56.7	13.6	13.1–14.9	5
9	159	48.4/51.6	14.6	14.1–15.4	4
10	235	38.7/61.3	15.8	15.3–16.9	6
11	154	43.5/56.5	16.8	16.3–17.5	4
12	81	40.7/59.3	17.8	17.3–19.0	2
All	807	40.6/59.4	15.5	13.1–19.0	21

potential classes provided by the principals of the participating schools. There are 21 intact classes involved in the final data for this paper. However, as noted above, students in the 12th grade have to prepare for the National High School Graduation Examination for university or college, so setting up a study programme is quite challenging. This study thus only assessed two classes, one in an urban area and one outside it. The students voluntarily joined the study after their teachers introduced the project. Most of the data were collected in March 2020, and some were collected in April 2021 (because of COVID-19 pandemic restrictions). The students completed the test in 45 min either in paper-and-pencil or online administration mode, depending on the particular conditions in each participating school. They took the online test via the eDia platform with a unique code (see Csapó & Molnár, 2019).

4.2. Instruments

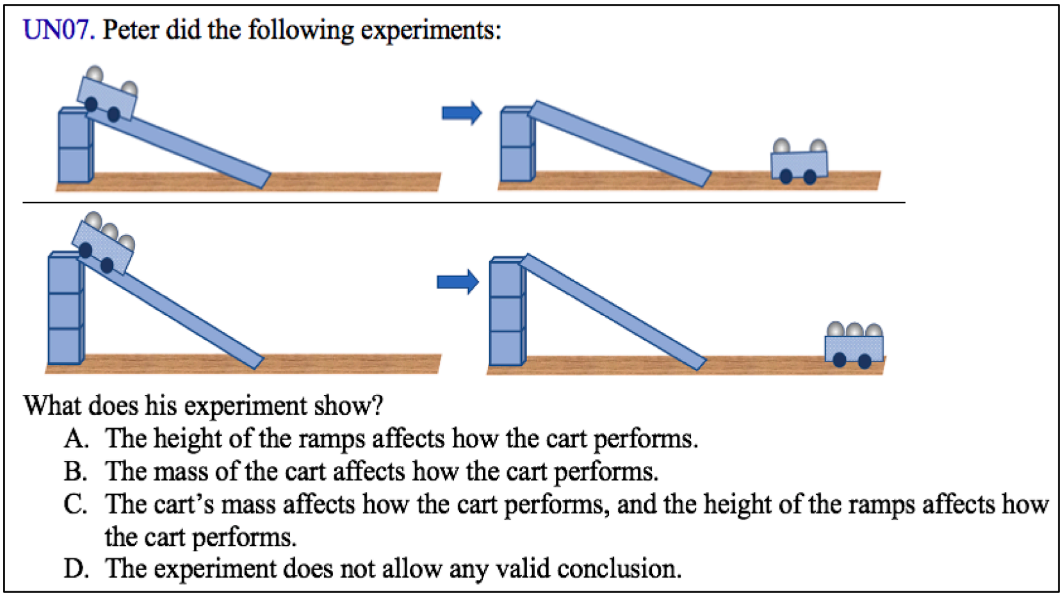
4.2.1. The control of variables strategy in physics (CVSP) test

The CVSP test consisted of 24 items in three CVS subskills: ID (eight items), IN (eight items), and UN (eight items). Due to the limitations of the multiple-choice format and the paper-based version, which are insufficient to measure the CVS subskill of PL (M. Schwichow et al., 2016), these subskill items were not included on the current CVSP test. Fig. 1 presents an item that falls under the UN subskill. The knowledge content of the test items related to basic physics in Vietnamese physics curricula for the secondary education level, such as mechanics, heat, thermodynamics, electricity, and electromagnetism. The test covered both cognitive processes in terms of the three subskills in CVS and domain content in physics.

The test items were presented in a multiple-choice format that contained a stem with three distractors and one correct answer. As it is the most popular format for testing scientific reasoning skills (Opitz et al., 2017), we used multiple-choice questions. Further, the opportunity for precision increases when measuring a larger number of respondents with smaller effect sizes than other question formats (Schwichow et al., 2016). We also endeavoured to minimize the impact of the students' reading ability levels by reducing the texts and using more visualized representations with figures, tables, and graphs. The reliability and validity of the test items have been confirmed in previous studies (Schwichow et al., 2020; Schwichow et al., 2016; Van Vo & Csapó, 2021b), and measurement invariance through differential item functioning analysis with respect to gender and administration mode has also been discussed in previous papers in the same context (Van Vo, 2022; Van Vo & Csapó, 2021b). Correct and incorrect answers were assigned a score of 1 and 0 points, respectively (see the CVSP test in Supplementary Material A).

4.2.2. Background questionnaire

The background questionnaire was adapted from PISA 2015 (OECD, 2017) to gather the students' background information, such as gender, grade level, and parents' educational attainment. On the self-report form, students were asked about the highest level of schooling their parents had completed, followed by a list of seven options for educational attainment ("Did not complete Grade 5," "Completed Grade 5," "Completed Grade 9," "Completed Grade 12," "Earned vocational or technical certificate/diploma," "Earned Bachelor's degree," and "Earned Master's degree or PhD"). The students were also asked to report the grade they had received on their final physics test in the previous semester. The tests are a kind of summative assessment and administered by each school with the approval of the local department of education. The background questionnaire is placed in the first section of the test instrument.



UN07. Peter did the following experiments:

What does his experiment show?

A. The height of the ramps affects how the cart performs.
B. The mass of the cart affects how the cart performs.
C. The cart's mass affects how the cart performs, and the height of the ramps affects how the cart performs.
D. The experiment does not allow any valid conclusion.

Fig. 1. A sample item for the understanding subskill.

4.3. Data analysis

4.3.1. Scaling the scores and reliability of the CVSP test

Rasch models are the most common psychometric approach in research on scientific reasoning (Edelsbrunner & Dablander, 2019). In this study, we employed a Rasch model measurement with the ACER ConQuest software (Adams & Wu, 2010; Wu et al., 2007). The One-Parameter Logistic Model and the marginal maximum likelihood estimator were used for dichotomous items. To test the model fit, we referred to the fit for single items (weighted mean squares, MNSQ) based on the cut-off standard with a range of 0.77–1.30 (Griffin, 2010).

The raw data were converted for the maximum likelihood estimation (MLE) scale as an output parameter in the Rasch model measurement. MLE is a statistical technique used to estimate the parameters of a probability distribution given some observed data (Wu et al., 2007). The standardization of the raw scores has been shown to be more accurate to present students' competencies (OECD, 2009). To explore the differences in the internal structure of the test, we fitted unidimensional and three-dimensional models to the dataset using the Rasch model. The infit indices ranged from 0.84 to 1.27, and the average value of the unidimensional model was 0.99 (SD = 0.10). The three-dimensional model had an infit range from 0.85 to 1.23 and a mean infit of 1.0 (SD = 0.1), indicating that the test showed a reliable construct and statistical fit in measuring CVS ability (see Appendix A). This will be instrumental for further analysis and interpretation of the results.

For the whole test, Cronbach's alpha was 0.81 and McDonald's omega (ω) was 0.82. The Cronbach's alpha values for the ID, IN, and UN subskills were 0.66 (ω = 0.70), 0.65 (ω = 0.68), and 0.55 (ω = 0.57), respectively. The internal consistency estimates were within acceptable levels (Taber, 2018). Pearson's �correlations ranged from 0.480 to 0.611, suggesting a significantly positive, high correlation between the three subskills. Table 2 presents intercorrelations between the subskills and understudied variables. All the included variables have a positive correlation, except for the gender variable. The gender variable only has a significant correlation with the UN subskill.

Item difficulty ranged from −1.43 to 2.75 in the unidimensional model and from −1.77 to 2.03 in the three-dimensional one. The average item difficulty (defaulted as 0 logits) compared to the average person proficiency of –0.08 showed that the students' proficiency was a little lower than the average item difficulty for the whole test. For single subskills, the average person proficiency in ID (+0.70), IN (−0.18), and UN (−0.77) suggested that the students were more proficient on the ID items and less proficient on the IN and UN items, among which the UN items seemed to be the most difficult for the test-takers.

4.3.2. Main data analysis strategy

Differential item functioning (DIF) analysis was used to inspect invariant measurement. This approach allows a comparison between different abilities among members of separate groups to ensure it is fair among groups at item level. We used the R difR package (Magis et al., 2010) for DIF analysis in the current study. Specifically, the Mantel–Haenszel (MH) method was employed to investigate equivalence for gender and grade level. The MH method uses indices, such as MH statistics or chi-square, *p*-value, and the effect size Δ_{MH} (ETS Delta). The cut-off criteria for MH statistics are 3.842 $p < 0.05$, and the effect size is as follows: $|\Delta_{MH}| \leq 1$ negligible, $1 < |\Delta_{MH}| \leq 1.5$ moderate, and $|\Delta_{MH}| > 1.5$ large (Zwick et al., 1999). An item with a negative effect size is likely to be more challenging for members of the focal group.

We visualized the distributions of the students' CVS capacity in different grade cohorts with the pirate plots in the yarrr package (Phillips, 2016). These plots include the mean, 95% shaded highest density intervals, jittered individual data points, and symmetric kernel densities; they thus offer more information than box plots and traditional bar plots (Phillips, 2017).

The developmental trend of the students' CVS proficiency was modelled with a nonlinear regression function. There is a variety of functional forms that can be parameterized to characterize a growth pattern (McNeish et al., 2020). In the study, the empirical data is modelled with a symmetric logistic function for model fitting similar to the one used in item response theory to obtain the curve (Hambleton & Swaminathan, 2013). This approach is frequently used to describe information for a developmental process in biology (Kniss & Streibig, 2018) and educational research (e.g., Ding, 2018; Han, 2013; Molnár et al., 2013). Generally, the logistic equation is as follows:

$$y = c + \frac{d - c}{1 + exp(b(log(x) - log(ED50)))}$$

Table 2
Intercorrelations (Pearson) between the subskills and understudied variables.

Variables	ID	IN	UN	GE	ME	FA
IN	.611***					
UN	.480***	.540***				
GE	–.012	–.031	–.083*			
ME	.194***	.154***	.159***	.001		
FA	.192***	.148***	.199***	.018	.673***	
PT	.276***	.253***	.186***	.039	.195***	.244***

Note: ID: identifying controlled experiments; IN: interpreting controlled experiments; UN: understanding the indeterminacy of confounded experiments; GE: gender; ME: mother's education; FA: father's education; PT: physics test ***$p < .001$.

where y is the response, c is the lower limit of the response (minimum asymptote), d is the upper limit (maximum asymptote), b is the slope around the point of inflection (hill slope), and ED50 is the response halfway between the upper and lower limits (a direct estimate of the point of inflection).

The parameters of a sigmoid curve were calculated with the R drc package (Ritz et al., 2015). �The observed data are fitted into the symmetric logistic models to estimate the cross-grade progression trend, with x representing student grade levels and y denoting students' proficiencies. In the logistic model, we computed the derivative of the logistic equation at grade level = ED50 (slope = $-b/(d-c)/4$e) to identify the grade level at which the most rapid change happened.

We employed multivariable regression models with Bayesian model averaging (BMA) (Raftery et al., 2020) to investigate the relevant predictors of the individual CVS capacity in physics. BMA is a technique used in statistical model selection and averaging, where instead of choosing a single "best" model, multiple models are considered, and their predictions are combined based on their posterior probabilities. BMA allows for a comprehensive analysis of multiple models, considering their individual strengths and weaknesses, and combining their predictions to obtain more robust and reliable estimates. Compared to the traditional approach, this method was considered a potentially major advance in terms of both its predictive and explanatory capabilities (Genell et al., 2010; Hair et al., 2010).

Using the BMA method, all possible models were initially assessed based on their model fit, as determined by the Bayesian information criterion (BIC), resulting in derived probabilities for each model. The posterior probability of the effect of each explanatory variable was calculated by taking the average of the posterior model probabilities for each model fit. The average mean and standard deviation for each regression coefficient were estimated using the weighted averaging of the coefficients for each individual model (see more Raftery et al., 2020).

There are several metrics or criteria can also be used to measure of model fit in BMA. For example, Raftery et al. (2020) referred BIC, posterior probability and evidence indices. The BIC index penalizes models that have a larger number of parameters, effectively balancing model fit and model complexity. Lower BIC values indicate better fit to the data while taking into account the number of parameters in the model. Posterior probability represents the updated belief in that model's validity or relevance after considering the available data. It takes into account both the prior probability assigned to the model and the likelihood of the data given that model. Models with higher posterior probabilities are considered more likely or better supported by the data, while models with lower probabilities are deemed less likely or less supported. Evidence index refers to the degree of support provided by the data for each model in the analysis, as measured by the posterior probabilities or Bayesian model weights. By comparing the evidence indices of different models, it is possible to determine which models are better supported by the data. Models with higher evidence indices are considered more likely to be the "true" model given the available evidence.

Other statistical tests, i.e., MANOVA, ANOVA, and the *t*-test, were used in the R program packages (R Core Team, 2019), such as psych (Revelle, 2019), and the graphs were drawn with ggplot2 (Wickham, 2016).

5. Results

5.1. *The patterns of students' performance in different grade cohorts*

DIF analysis was employed to examine statistically invariant characteristics at the item level. We conducted the measurement invariance test with DIF analysis as a prerequisite to ensure the same measure construct across grade levels. After we used the MH method with the 10th grade (the middle grade cohort) as a focal group and others as reference groups in the difR package (Magis et al., 2010), the results suggested two DIF items (ID06 and IN06), in which IN06 favoured the 10th graders and ID06 was more difficult for

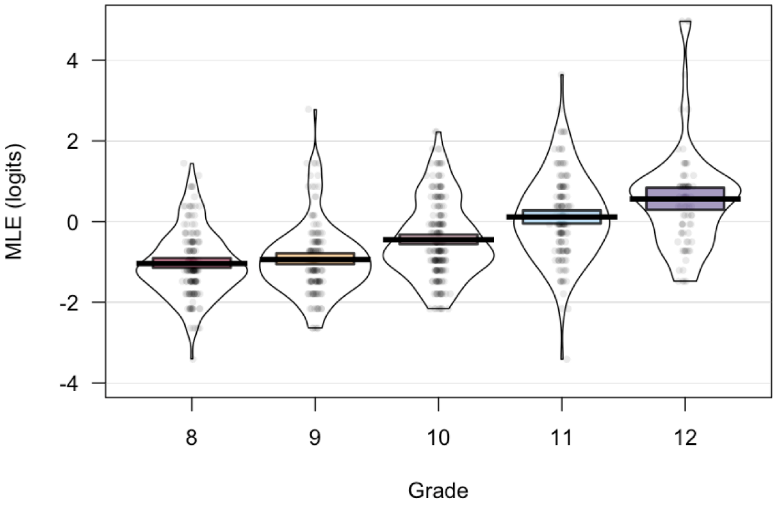


Fig. 2. Differences in performance of grade cohorts on the CVS test.

them than others (see more results of the DIF analysis in Appendix B). These items were thus excluded in further relevant analysis.

The students' performance on the CVSP tests across grade levels is illustrated in the pirate plots (Fig. 2). In general, the students in the older groups performed better than their juniors, and the score distributions differed between the grade cohorts. The mean score of the 8th graders was the lowest ($M = -1.04$, SD = 0.82), with the majority of the students receiving a low score, while the 9th graders performed slightly better ($M = -0.94$, SD = 0.89), with several students attaining very high scores. Distribution in the 10th graders' scores was similar to that of the 8th grade, but its score was notably higher on average ($M = -0.45$, SD = 0.95). The mean scores in the 11th- and 12th-grade groups were 0.12 and 0.56, respectively, but their score distributions displayed different shapes. In the 11th grade, there were both high- and low-performing participants, and the proportion of students was equivalent. However, several 11th-graders achieved high scores on the test. The distribution shape tended to spread up to the top of the scale in the 12th-grade cohort (see distributions of the raw scores in Supplementary Material B).

Furthermore, MANOVA analysis was used to examine the effect of grade levels on students' CVSP competence in the three subskills. The results demonstrated that there were significant discrepancies between the school grade groups on the three subtests [F(12, 2406) = 18.20, $p < .001$]. Additionally, ANOVA analysis showed that a significant difference was indicated on the students' CVSP proficiency in general [F(4, 802) = 61.31, $p < .001$]. Follow-up univariate ANOVAs with a Bonferroni–Holm adjusted alpha level of 0.0167 (Barbara & Linda, 2012) indicate that there were significant discrepancies between the school grade groups on the subtests: ID [F(4, 802) = 41.89, $p < .001$]; IN [F(4, 802) = 40.63, $p < 0.001$]; and UN [F(4, 802) = 32.92, $p < 0.001$]. Additionally, we employed Tukey's HSD analysis to identify the differences in pairs of grades (Table 3). The results demonstrate significant differences in all the grade pairs, except the 9th–8th grades. Overall, the findings suggest that students from the higher grade levels performed better than their counterparts in the younger groups.

We used the four-parameter symmetric logistic function (Ritz et al., 2015) to model the empirical data and visualize the developmental process with smooth lines (Wickham, 2016). As shown in Fig. 3, the fitted logistic curve was presented by the smooth line, since the dots represented the empirical mean scores. The approximate F-test (see Kniss & Streibig, 2018) was employed to check the model fit by comparing the model to the ANOVA model. The results showed that the logistic curve fit quite well with the empirical data (F=0.519, $p = .471$). The Akaike information criterion (AIC) was used to check whether the nonlinear model is better or worse than the linear model. The results indicate that the nonlinear model (AIC = 2249.03) was a better fit to the data than the linear one (AIC = 2256.53). This means that the development in students' CVS capacity in physics was significant and nonlinear across the grade cohorts but that the growth rate was different. The most significant growth was identified around the 10th grade as �the point of inflexion (ED50 – half maximal effective concentration). This suggests that the fastest development occurred from the 10th grade to the 11th grade, with a score of approximately 0.56 (digits) per year. The actual slope of the tangent of the curve is around 2.70 (or 69.7°).

The developmental patterns of the individual subskills were modelled into a four-parameter symmetric logistic equation as depicted in Fig. 4. The logistic curves demonstrated that the models fitted well with the empirical data (ID: F=0.493; $p = .483$; IN: F=0.918, $p = .339$; and UN: F=0.343, $p = .558$). Students across grade levels achieved the highest scores on the ID subskill, followed by the IN and the UN items, the latter of which seemed to be difficult for the students. The changing patterns of students' performance on the ID and IN subskills appeared to be similar across the grade levels, with the most rapid change from the 10th to 11th grades.

The results showed that the progression trends for students' IN ability had a better fit to the nonlinear model (AIC = 2745.42), compared to the linear one (AIC = 2798.91). Likewise, the cross-grade progression trend of the students' performance on the IN subskill fitted better in the nonlinear model (AIC = 2745.42) than the linear model (AIC = 2770.90).

On the UN subskill, however, the developmental curve for the students' scores fitted better to the linear model (AIC = 2538.79) compared to the understudied nonlinear one (AIC = 2745.42), suggesting the students' performance on the UN subskill improved linearly through the grade cohorts.

5.2. Gender difference in CVSP

We conducted a DIF analysis to determine if the items on the test were biased against subsamples on the basis of gender. Overall, the results of the DIF with the MH method (the female students as the focal group and males as the reference group) showed that two items (items IN08 and UN04) were flagged as DIF items. Item IN08 favoured the female students, while item UN04 favoured the males, with a moderate effect in both. Therefore, these two items were excluded from the relevant analysis. Item UN07 was flagged at the B level (moderate effect), but it was not detected as a DIF item because its p-value was greater than 0.05. Interestingly, a positive effect size was found for twelve items, indicating that they favoured female students, while the remaining twelve items had negative effect size

Table 3
Tukey's HSD: multiple comparisons.

Grades	ID		IN		UN		CVSP	
	Mean difference	p	Mean difference	p	Mean difference	p	Mean difference	p
9th & 8th	0.049	.997	0.019	.999	0.280	.177	0.100	.883
10th & 9th	0.672	<0.001	0.572	<0.001	0.35	.032	0.492	<0.001
11th & 10th	0.644	<0.001	0.682	<0.001	0.47	.001	0.563	<0.001
12th & 11th	0.279	.537	0.484	.066	0.36	.168	0.445	.008

Note: ID: identifying controlled experiments; IN: interpreting controlled experiments; UN: understanding the indeterminacy of confounded experiments; CVSP: control of variables strategy in physics test.

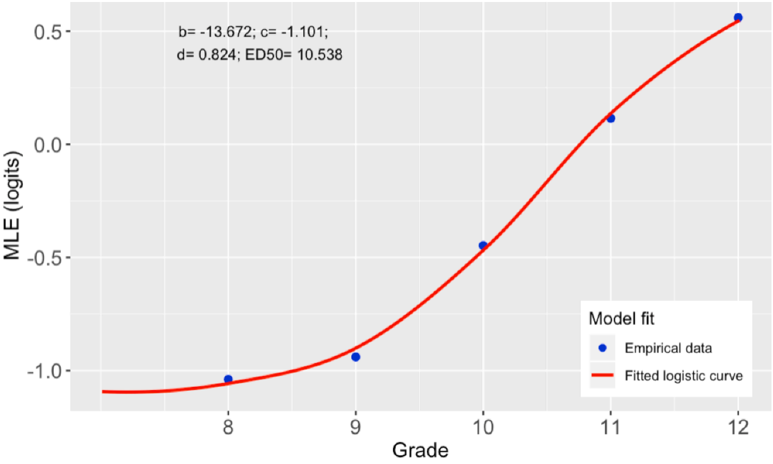


Fig. 3. The developmental trend of CVS in physics.

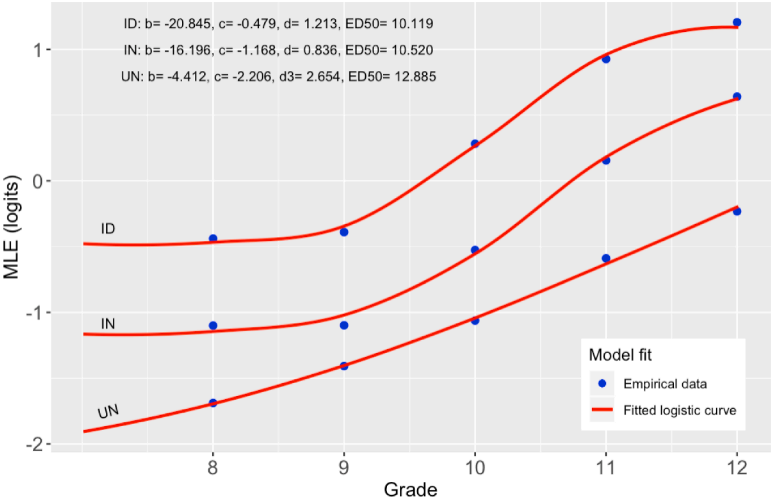


Fig. 4. The developmental curves for each CVS subskill.
Note: ID: identifying controlled experiments; IN: interpreting controlled experiments; UN: understanding the indeterminacy of confounded experiments

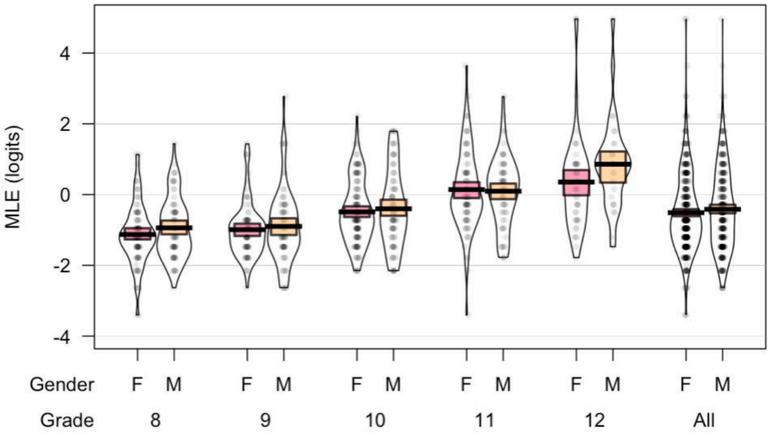


Fig. 5. Comparison of performance on the CVSP test between males and females.

values, suggesting that they favoured male students (see more results of the DIF analysis in Appendix C).

As demonstrated in Fig. 5, both the male and female students seemed to perform equivalently on the CVSP test in the single-grade cohorts and the whole sample in general. However, the male students' average score (Mean = −0.42, SD = 1.17) was a little higher than that of the female students (Mean = −0.51, SD = 1.03) in general. The distributions displayed slight differences between the boys and girls, with the larger part of the density spread over a wider range in the boy's groups.

We employed the *t*-test to compare abilities between the male and female students. Table 4 summarizes the results of the *t*-test for single groups and the whole sample. No significant differences were found between the boys and girls on the CVSP test in either the individual cohorts or the entire sample, suggesting that the boys did not differ significantly from the girls on the CVSP test.

Furthermore, Table 5 provides the results of the *t*-test for each subskill by gender. The results show that no gender difference was found for ID, IN, and UN at most grade levels, except in the 10th grade, where the male students performed significantly better than their female classmates for the UN subskill.

5.3. *Predicting the students' CVSP*

In order to predict the students' CVS capacity in physics, the exploratory variables included grade level (or student age), gender, the content knowledge test (results of the physics test in the previous semester), mother's educational attainment, and father's educational attainment. We conducted the BMA analysis to explore possible models in predicting students' CVSP based on the understudied variables with linear regression models in R package BMA (Raftery et al., 2020). The best four models were recommended based on the Bayesian information criterion as presented in Table 6. Model 1 was suggested as the best one, which included three variables (i.e., grade level, physics test, and mother's education), since it had the highest posterior model probability (62.10%) and the lowest BIC value (−285.198) and can explain around 34.3% of predicted variance. The next best model is model 2, which consisted of four variables: grade level, physics test, mother's education, and father's education. This model can explain 34.6% of variance with a posterior probability of 13.9 per cent. Following the three-factor model, it included the grade level, physics test, and father's education variables and can explain 34.0% variance, but its posterior probability was 13.7 per cent. The predictors in model 4 were similar to those in model 1 with one additional variable: gender. Although this model can explain 34.6% variance of CVSP, its posterior probability was only 10.3 per cent.

6. Discussion and conclusions

The cross-grade progression trend in students' CVSP scores showed a significant increase across the study cohorts. The research employed symmetric logistic functions to simulate the developmental curves, thereby yielding further insights into the CVS subskills at various grade levels, i.e., flooring and ceiling baselines, and average variations due to population differences. Generally, these results partly correspond with those of previous studies (e.g., Bullock & Ziegler, 1999; Schwichow et al., 2020; Tairab, 2015), but the fastest period of the developments appeared to occur slightly later, according to recent findings. Bullock and Ziegler (1999) discovered that ID progresses in most children during the third and fourth grades. Meanwhile, Schwichow et al. (2020) observed that students' scores on the ID and IN subtests improved throughout their lower secondary school years, spanning from the fifth to the ninth grades. Our investigation showed that the cross-grade progression curve of the ID and IN subtests had a direct estimate for the point of inflection during the early and later tenth grade. There may be various reasons for the inconsistent results. The reasons may stem from the effects of current physics curricula and educational practices. When schools teach thinking skills in embedded programmes, the development of skills occurs spontaneously as a 'by-product' of teaching ordinary school material rather than through explicit instruction (de Koning, 2000). This suggests that teachers in the lower secondary grades need more awareness of how to incorporate CVS problem tasks into school practice. Other potential reasons for why the 9th graders were less proficient in CVSP concern the primary attention given to the provincial selection examination for public high school, which involves the three core school subjects, mathematics, literature, and English. Nevertheless, the finding is consistent with those of Han (2013) in the context of China. Further, it is in keeping with Chinese research (Ding, 2018; Han, 2013), in which it was argued that teaching physics in high school appeared to be more focused on content knowledge and lack of inquiry activities to promote students' CVS skills.

As regards a potential gender discrepancy, both male and female students performed equivalently on the CVSP test. These results are in line with Mayer et al. (2014), Piraksa et al. (2014), and Thuneberg et al. (2015), but they do not agree with other studies (e.g., Tairab, 2015; Tekkaya & Yenilmez, 2006; Valanides, 1997). The findings may be linked to the Vietnamese context, where more young females tend to venture into science-related sectors than in the past (International Labor Organization, 2020). However, a statistically

Table 4
The *t*-test comparing the CVSP test results by gender.

	Male		Female			
Grade	N	*Mean (SD)*	N	*Mean (SD)*	t	p
8	77	−0.94 (0.85)	101	−1.13 (0.85)	1.46	.147
9	77	−0.90 (1.03)	82	−0.99 (0.78)	0.63	.527
10	99	−0.40 (1.10)	144	−0.49 (0.90)	0.65	.520
11	67	0.15 (0.98)	87	0.15 (1.13)	−0.02	.987
12	33	0.85 (1.25)	48	0.36 (1.35)	1.72	.090
All	345	−0.42 (1.17)	462	−0.51 (1.10)	1.18	.238

Table 5
The *t*-test comparing the results between males and females on each subtest.

Grade	ID				IN				UN			
	M_b	M_g	t	p	M_b	M_g	t	p	M_b	M_g	t	p
8	−0.34	−0.50	0.79	.428	−1.19	−1.46	1.35	.178	−1.54	−1.70	0.92	.363
9	−0.22	−0.51	1.36	.177	−1.22	−1.45	1.10	.275	−1.47	−1.30	−0.90	.369
10	0.23	0.30	−0.37	.715	−0.70	−0.77	0.34	.737	−0.81	−1.19	2.43	.016
11	0.89	1.07	−0.76	.446	−0.07	−0.18	0.49	.628	−0.57	−0.58	0.27	.790
12	1.57	1.14	1.64	.105	0.59	−0.01	1.93	.058	0.18	−0.34	1.51	.137
All	0.26	0.22	0.43	.667	−0.68	−0.85	1.61	.108	−0.97	−1.12	1.57	.116

Note: M_b: mean score for boys; M_g: mean score for girls.

Table 6
The best four models selected by BMA analysis to predict students’ CVSP.

Variable	p!= 0	EV	SD	Model 1	Model 2	Model 3	Model 4
Grade level	100.0	0.438	0.027	0.439	0.436	0.433	0.439
Gender	10.3	−0.012	0.042				−0.119
Physics test	100.0	0.167	0.020	0.168	0.162	0.164	0.170
Mother’s education	86.3	0.089	0.043	0.109	0.074		0.108
Father’s education	27.6	0.021	0.038		0.052	0.099	
Number of variables				3	4	3	4
R-square				0.343	0.346	0.340	0.346
BIC				−285.198	−282.200	−282.177	−281.611
Posterior probability				0.621	0.139	0.137	0.103

Note: EV: evidence index.

significant difference was observed in the UN subskill, which favoured the male students. Based on these results, the effects of a gender disparity in the Vietnamese context should be studied in more detail.

The BMA analysis suggested a possible model for explaining individual CVSP among students based on the grade level (or student age), content knowledge test, and parents’ educational attainment variables. In the better model, the grade level was shown as the best factor to predict CVSP in children, followed by the physics test (content knowledge) and mother’s educational attainment. The finding confirmed that content knowledge and scientific reasoning are interrelated, as content knowledge provides a foundation for scientific reasoning and understanding. This outcome is in line with previous research (e.g., Bao et al., 2009; Coletta & Phillips, 2005;

Appendix A
Summary of psychometric characteristics of items on the CVSP test.

No.	Item	Subskill	Correct answer (%)	Dis	One-dimensional		Three-dimensional	
					Dif	Infit	Dif	Infit
1	ID01	ID	59.36	0.44	−0.94	1.01	−0.27	1.05
2	ID03	ID	33.46	0.30	0.33	1.14	1.08	1.17
3	IN02	IN	28.00	0.21	0.64	1.17	0.59	1.23
4	ID02	ID	56.01	0.52	−0.78	0.93	−0.10	0.97
5	ID04	ID	68.77	0.40	−1.43	1.00	−0.78	1.03
6	UN08	UN	41.26	0.28	−0.07	1.14	−0.65	1.09
7	IN05	IN	31.97	0.40	0.41	1.04	0.35	1.05
8	IN07	IN	38.29	0.50	0.08	0.96	0.01	0.99
9	ID05	ID	65.43	0.54	−1.25	0.91	−0.59	0.89
10	UN02	UN	50.68	0.55	−0.53	0.93	−1.08	0.89
11	ID06	ID	56.88	0.62	−0.82	0.84	−0.14	0.85
12	IN04	IN	33.21	0.55	0.34	0.91	0.28	0.93
13	UN03	UN	7.56	0.17	2.42	1.02	1.71	0.99
14	ID07	ID	46.22	0.40	−0.31	1.02	0.40	1.11
15	IN01	IN	42.38	0.44	−0.13	1.03	−0.20	1.04
16	IN03	IN	49.19	0.60	−0.45	0.87	−0.54	0.90
17	UN06	UN	20.45	0.55	1.13	0.86	0.47	0.86
18	IN08	IN	60.59	0.54	−1.01	0.91	−1.11	0.96
19	UN07	UN	5.70	0.17	2.75	0.98	2.03	0.99
20	UN01	UN	46.96	0.45	−0.35	1.00	−0.91	0.96
21	UN04	UN	24.54	0.40	0.85	1.02	0.21	0.95
22	ID08	ID	46.72	0.52	−0.31	0.93	0.40	0.95
23	UN05	UN	65.55	0.06	−1.26	1.27	−1.77	1.17
24	IN06	IN	27.26	0.48	0.68	0.92	0.63	0.96

Note: ID: identifying controlled experiments; IN: interpreting controlled experiments; UN: understanding the indeterminacy of confounded experiments; Dis: discrimination; Dif: item difficulty.

Appendix B
Results of DIF analysis with the MH method by grade level.

Order	Item	MHχ^2	p	α_{MH}	Δ_{MH}	Label
1	ID01	0.133	.716	0.924	0.187	A
2	ID03	1.593	.207	0.796	0.537	A
3	IN02	0.947	.331	0.832	0.431	A
4	ID02	0.771	.380	0.840	0.410	A
5	ID04	0.266	.606	0.893	0.266	A
6	UN08	0.064	.800	1.058	−0.133	A
7	IN05	1.323	.250	1.261	−0.544	A
8	IN07	0.008	.929	0.999	0.001	A
9	ID05	0.675	.411	1.199	−0.427	A
10	UN02	0.889	.346	0.828	0.444	A
11	ID06	4.545	.033	1.617	−1.130	B
12	IN04	0.033	.857	0.946	0.130	A
13	UN03	0.659	.417	1.410	−0.807	A
14	ID07	0.163	.686	1.087	−0.197	A
15	IN01	2.263	.133	1.329	−0.668	A
16	IN03	0.433	.510	1.157	−0.342	A
17	UN06	1.315	.252	0.748	0.683	A
18	IN08	0.004	.952	0.970	0.072	A
19	UN07	1.821	.177	0.581	1.276	B
20	UN01	0.003	.954	1.026	−0.061	A
21	UN04	0.582	.446	0.845	0.397	A
22	ID08	1.533	.216	1.280	−0.580	A
23	UN05	0.812	.368	1.177	−0.384	A
24	IN06	6.001	.014	0.595	1.220	B

Note: effect size code: A: negligible effect; B: moderate effect; C: large effect; detection threshold: 3.842; focal group: 10th grade. Items detected as DIF items: ID06 and IN06.

Appendix C
Results of DIF analysis with the MH method by gender.

ORDER	ITEM	MHχ^2	p	α_{MH}	Δ_{MH}	LABEL
1	ID01	1.566	.211	0.802	0.519	A
2	ID03	0.547	.460	0.877	0.309	A
3	IN02	1.877	.171	0.781	0.582	A
4	ID02	2.462	.117	0.751	0.672	A
5	ID04	0.090	.764	1.070	−0.159	A
6	UN08	0.061	.806	0.952	0.117	A
7	IN05	0.565	.452	0.866	0.338	A
8	IN07	0.001	.973	0.991	0.021	A
9	ID05	0.002	.966	0.975	0.061	A
10	UN02	0.002	.967	0.978	0.053	A
11	ID06	0.970	.325	1.231	−0.487	A
12	IN04	1.039	.308	1.233	−0.493	A
13	UN03	3.298	.069	1.793	−1.372	B
14	ID07	0.401	.527	0.893	0.266	A
15	IN01	0.007	.932	1.028	−0.064	A
16	IN03	2.915	.088	1.394	−0.780	A
17	UN06	3.742	.053	1.551	−1.031	B
18	**IN08**	**8.080**	**.005**	**0.591**	**1.238**	**B**
19	UN07	2.179	.140	1.799	−1.380	B
20	UN01	0.826	.363	1.172	−0.372	A
21	**UN04**	**5.541**	**.019**	**1.562**	**−1.047**	**B**
22	ID08	0.500	.479	1.146	−0.320	A
23	UN05	2.975	.085	0.760	0.645	A
24	IN06	0.394	.530	1.147	−0.322	A

Note: effect size code: A: negligible effect; B: moderate effect; C: large effect; detection threshold: 3.842; focal group: female. Items detected as DIF items: IN08 and UN04.

Schwichow et al., 2020; Song & Black, 1992; Van Vo & Csapó, 2021b). Notably, the father's education variable was not involved in the best model in predicting students' CVSP, but it is still considered among the possible models for this purpose. The culture is a potential reason because parents in Vietnam, especially mothers, take a particularly strong interest in their children's performance in school (Hoang et al., 2014; Phan, 2004). In addition, the influence of parents' education on their children's scientific reasoning was also partly reflected in Spruijt et al. (2020). This is quite consistent with previous studies, which found a significant influence of parental education on school achievement (Gienger, Petermann, & Petermann, 2008) and scientific reasoning (Koerber et al., 2015; Van Vo & Csapó, 2021a, 2023).

The findings demonstrated that the students felt the item bundle for the UN subskill was more difficult than those for the ID and IN subskills across grade levels. This suggests that when using CVS tasks in the teaching process, teachers should provide easy ones (ID and IN) rather than more difficult ones (UN). As Siler and Klahr (2012) indicate, students often misunderstand the goal of a task as contriving an outcome rather than ascertaining the causal status of a single variable. Teachers should therefore help students to distinguish between confounded experiments and controlled experiments on UN tasks. They can use invalid designs as effective examples to teach the logic of CVS (e.g., Chen & Klahr, 1999; Lorch et al., 2014). Students often confuse non-influential variables and non-testable variables (Zhou et al., 2016), thinking that they can 'do it all' in a single experiment (Siler & Klahr, 2012); consequently, they cannot distinguish between confounded experiments and controlled experiments. Teachers should assist them by raising awareness to these possible traps in completing the CVS tasks. Additionally, Bayesian inference was demonstrated as a potential approach in educational research. In particular, BMA analysis can be replicated in order to evaluate appropriate models in educational studies.

This study has some limitations. The findings were drawn from cross-sectional data that may be biased due to the influence of different environments (Maxwell & Cole, 2007). Future studies should consider a longitudinal approach to examine the success of curricula through multiple-administration assessment, wherein a student is repeatedly tested over time with dynamic measurement models (e.g., McNeish et al., 2020). The current test did not include the PL task and experimental data, which assess an in-depth interpretation beyond the testability of the variables in an experiment (Han, 2013; Zhou et al., 2016). The results of the physics tests may be quite biased because they were collected before and during the pandemic, and such conditions may influence students' performance. Therefore, it is prudent to be cautious when drawing general conclusions about content knowledge and CVSP. These limitations should be considered in future research.

CRediT authorship contribution statement

De Van Vo: Conceptualization, Writing – original draft, Investigation, Methodology, Visualization, Software, Data curation, Writing – review & editing. **Benő Csapó:** Supervision, Validation, Methodology, Investigation, Data curation, Writing – review & editing. **Samuel Greiff:** Methodology, Writing – review & editing.

Declaration of Competing Interest

The authors have stated no potential conflict of interest.

Data availability

Data will be made available on request.

Acknowledgments

The authors are thankful to the student participants and the teachers who assisted with the assessments in the secondary schools. The Center for Research on Learning and Instruction at the University of Szeged supported the online administration to the Vietnamese students via the eDia platform.

Funding

Open access funding provided by University of Szeged [6299].

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.tsc.2023.101371.

References

Adams, R., & Wu, M. (2010). Modelling a dichotomously scored multiple choice test with the Rasch model (Issue August). *ConQuest*.

Adey, P., Csapó, B., Csapó, B., & Szabó, G. (2012). Developing and assessing scientific reasoning. *Framework for diagnostic assessment of science* (pp. 17–53). Nemzeti Tankönyvkiadó.

Bao, L., Fang, K., Cai, T., Wang, J., Yang, L., Cui, L., et al. (2009). Learning of content knowledge and development of scientific reasoning ability: A cross culture comparison. *American Journal of Physics, 77*(12), 1118–1123. https://doi.org/10.1119/1.2976334

Barbara, G. T., & Linda, S. F. (2012). *Using multivarite statistics* (6th ed.). Pearson.

Becker, M., McElvany, N., & Kortenbruck, M. (2010). Intrinsic and extrinsic reading motivation as predictors of reading literacy: A longitudinal study. *Journal of Educational Psychology, 102*(4), 773–785. https://doi.org/10.1037/a0020084

Boudreaux, A., Shaffer, P. S., Heron, P. R. L., & McDermott, L. C. (2008). Student understanding of control of variables: Deciding whether or not a variable influences the behavior of a system. *American Journal of Physics, 76*(2), 163–170. https://doi.org/10.1119/1.2805235

Bullock M., & Ziegler A. (1999). Scientific Reasoning: Developmental and individual differences individual development from 3 to 12: Findings from the Munich longitudinal study. Cambridge University Press.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*(5), 1098–1120. https://doi.org/10.1111/1467-8624.00081

Coletta, V. P., & Phillips, J. A. (2005). Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability. *In American Journal of Physics, 73*(12), 1172–1182. https://doi.org/10.1119/1.2117109

Csapó, B., & Molnár, G. (2019). Online diagnostic assessment in support of personalized teaching and learning: The eDia system. *Frontiers in Psychology, 10*(JULY). https://doi.org/10.3389/fpsyg.2019.01522

Csapó, B. (1997). The development of inductive reasoning: Cross-sectional assessments in an educational context. *International Journal of Behavioral Development, 20* (4), 609–626. https://doi.org/10.1080/016502597385081

de Koning E. (2000). Inductive reasoning in primary education: Measurement, teaching, transfer. Unpublished doctoral dissertation. Tilburg University, Utrecht.

Ding, L. (2018). Progression trend of scientific reasoning from elementary school to university: A large-scale cross-grade survey among Chinese students. *International Journal of Science and Mathematics Education, 16*(8), 1479–1498. https://doi.org/10.1007/s10763-017-9844-0

Edelsbrunner, P. A., & Dablander, F. (2019). The psychometric modeling of scientific reasoning: A review and recommendations for future avenues. *Educational Psychology Review, 31*(1), 1–34. https://doi.org/10.1007/s10648-018-9455-5

Edelsbrunner, P. A., Schalk, L., Schumacher, R., & Stern, E. (2018). Variable control and conceptual change: A large-scale quantitative study in elementary school. *Learning and Individual Differences, 66*(November 2016), 38–53. https://doi.org/10.1016/j.lindif.2018.02.003

Edelsbrunner, P. A., Schumacher, R., Stern, E., Houdé, O., & Borst, G. (2022). The Cambridge handbook of cognitive development. *The cambridge handbook of cognitive development*. Cambridge University Press. https://doi.org/10.1017/9781108399838

Genell, A., Nemes, S., Steineck, G., & Dickman, P. W. (2010). Model selection in medical research: A simulation study comparing Bayesian model averaging and stepwise regression. *BMC Medical Research Methodology, 10*(1), 108. https://doi.org/10.1186/1471-2288-10-108

Gienger, C., Petermann, F., & Petermann, U. (2008). Wie stark hängen die HAWIK-IV-Befunde vom Bildungsstand der Eltern ab? *Kindheit und Entwicklung, 17*(2), 90–98.

Griffin, P. (2010). *Item response modelling: An introduction to the rasch model*. Assessment Research Centre Faculty of Education, The University of Melbourne.

Hair, J. F., William, J., Barry, C. B., Rolph, J. B., & Anderson, E. (2010). *Multivariate data analysis*. Pearson.

Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. Springer Science & Business Media.

Han, J. (2013). *Scientific reasoning: Research, development, and assessment (Unpublished doctoral dissertation)*. The Ohio State University.

Hejnová, E., Eisenmann, P., Cihlář, J., & Přibyl, J. (2018). Relations between scientific reasoning and culture of problem solving. *Journal on Efficiency and Responsibility in Education and Science, 11*(2), 38–44. https://doi.org/10.7160/eriesj.2018.110203

Hoang, K. M., Nguyen, H. T., & La, T. T. (2014). Parent and teacher communication: A case study in Vietnam. W. J., X. B., & W. B.. *Innovative management in information and production* (pp. 305–313) New York: Springer. https://doi.org/10.1007/978-1-4614-4857-0_33

Ifenthaler, D., & Seel, N. M. (2011). A longitudinal perspective on inductive reasoning tasks. Illuminating the probability of change. *Learning and Instruction, 21*(4), 538–549. https://doi.org/10.1016/j.learninstruc.2010.08.004

International Labour Organization. (2020). Leading to success: The business case for women in business and management in Viet Nam. https://ilo.org/hanoi/Whatwedo/Publications/WCMS_761063/lang–en/index.htm.

Kniss, A. R., & Streibig, J. C.. Statistical analysis of agricultural experiments using R. https://rstats4ag.org.

Koerber, S., Mayer, D., Osterhaus, C., Schwippert, K., & Sodian, B. (2015). The development of scientific thinking in elementary school: A comprehensive inventory. *Child Development,, 86*(1), 327–336. https://doi.org/10.1111/cdev.12298

Kuhn, D., & Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science, 16*(11), 866–870. https://doi.org/10.1111/j.1467-9280.2005.01628.x

Kuhn, D. (2007). Reasoning about multiple variables: Control of variables is not the only challenge. *Science Education, 91*(5), 710–726. https://doi.org/10.1002/sce.20214

Kwon, Y. J., & Lawson, A. E. (2000). Linking brain dowth with the development of scientific reasoning ability and conceptual change during adolescence. *Journal of Research in Science Teaching, 37*(1), 44–62. https://doi.org/10.1002/(SICI)1098-2736(200001)37:1<44::AID-TEA4>3.0.CO;2-J

Lawson, A. (2009). Basic inferences of scientific reasoning, argumentation, and discovery. *Science Education, 94*(2), 336–364. https://doi.org/10.1002/sce.20357

Lazonder, A. W., Janssen, N., Gijlers, H., & Walraven, A. (2021). Patterns of development in children\'s scientific reasoning: results from a three-year longitudinal study. *Journal of Cognition and Development, 22*(1), 108–124. https://doi.org/10.1080/15248372.2020.1814293

Lorch, R. F., Lorch, E. P., Freer, B. D., Dunlap, E. E., Hodell, E. C., & Calderhead, W. J. (2014). Using valid and invalid experimental designs to teach the control of variables strategy in higher and lower achieving classrooms. *Journal of Educational Psychology, 106*(1), 18–35. https://doi.org/10.1037/a0034375

Lorch, R. F., Lorch, E. P., Wheeler, S. L., Freer, B. D., Dunlap, E., Reeder, E. C., et al. (2020). Oversimplifying teaching of the control of variables strategy. *Psicologia Educativa, 26*(1), 7–16. https://doi.org/10.5093/PSED2019A13

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*, Article 100899. https://doi.org/10.1016/j.tsc.2021.100899

MOET. (2009). Tài liệu phân phối chương trình Vật lí THPT.Available online: https://hoatieu.vn/bieu-mau/phan-phoi-chuong-trinh-mon-vat-ly-bac-thpt-128692 (accessed on 20 May 2020).

Magis, D., Beland, S., Tuerlinckx, F., & Boeck, P. De (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847–862. https://doi.org/10.3758/BRM.42.3.847

Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods, 12*(1), 23–44. https://doi.org/10.1037/1082-989X.12.1.23

Mayer, D., Sodian, B., Koerber, S., & Schwippert, K. (2014). Scientific reasoning in elementary school children: Assessment and relations with cognitive abilities. *Learning and Instruction, 29*, 43–55. https://doi.org/10.1016/j.learninstruc.2013.07.005

McNeish, D., Dumas, D. G., & Grimm, K. J. (2020). Estimating new quantities from longitudinal test scores to improve forecasts of future performance. *Multivariate Behavioral Research, 55*(6), 894–909. https://doi.org/10.1080/00273171.2019.1691484

Molnár, G., Greiff, S., & Csapó, B. (2013). Inductive reasoning, domain specific and complex problem solving: Relations and development. *Thinking Skills and Creativity, 9*, 35–45. https://doi.org/10.1016/j.tsc.2013.03.002

OECD. (2009). The Rasch Model. *PISA data analysis manual: SAS* (2nd Ed., pp. 79–94). Paris: OECD Publishing. https://doi.org/10.1787/9789264056251-6-en

OECD. (2017). *PISA 2015 results students' well-being*. OECD. https://doi.org/10.1787/9789264273856-en. Vol. III.

Opitz, A., Heene, M., & Fischer, F. (2017). Measuring scientific reasoning – a review of test instruments. *Educational Research and Evaluation, 23*(3–4), 78–101. https://doi.org/10.1080/13803611.2017.1338586

Osterhaus, C., Koerber, S., & Sodian, B. (2017). Scientific thinking in elementary school: Children\'s social cognition and their epistemological understanding promote experimentation skills. *Developmental Psychology, 53*(3), 450–462. https://doi.org/10.1037/dev0000260

Phan, T. (2004). A qualitative study of Vietnamese parental involvement and their high academic achieving children. *Journal of Authentic Learning, 1*, 51–61.

Phillips, N. (2016). Yarrr !. *The pirate 's guide to R*. https://bookdown.org/ndphillips/YaRrr/.

Phillips, N. (2017). *Yarrr: A companion to the e-Book "YaRrr!: The pirate's guide to R*. https://bookdown.org/ndphillips/YaRrr/.

Piekny, J., Grube, D., & Maehler, C. (2014). The development of experimentation and evidence evaluation skills at preschool age. *International Journal of Science Education, 36*(2), 334–354. https://doi.org/10.1080/09500693.2013.776192

Piraksa, C., Srisawasdi, N., & Koul, R. (2014). Effect of gender on student\'s scientific reasoning ability: A case study in Thailand. *Procedia - Social and Behavioral Sciences, 116*, 486–491. https://doi.org/10.1016/j.sbspro.2014.01.245

R Core Team. (2019). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. https://www.r-project.org/.

Raftery A., Hoeting J., Volinsky C., Painter I., Yeung K.Y. (.2020). BMA: Bayesian model averaging. Available online: https://cran.r-project.org/web/packages/BMA/BMA.pdf (accessed on 16 December 2022).

Revelle, W. (2019). *psych: Procedures for psychological, psychometric, and personality research*. Northwestern University. https://cran.r-project.org/package=psych.

Ritz, C., Baty, F., Streibig, J. C., & Gerhard, D. (2015). Package "drc": Analysis of Dose-Response Curves. *PloS One, 10*(12), Article e0146021. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0146021.

Schauble, L., Klopfer, L. E., & Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching, 28*(9), 859–882. https://doi.org/10.1002/tea.3660280910

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology, 32*(1), 102–119. https://doi.org/10.1037/0012-1649.32.1.102

Schlatter, E., Molenaar, I., & Lazonder, A. W. (2020). Individual differences in children\'s development of scientific reasoning through inquiry-based instruction: who needs additional guidance? *Frontiers in Psychology, 11*, 1–14. https://doi.org/10.3389/fpsyg.2020.00904

Schwichow, M., Croker, S., Zimmerman, C., Höffler, T., & Härtig, H. (2016a). Teaching the control-of-variables strategy: A meta-analysis. *Developmental Review, 39*, 37–63. https://doi.org/10.1016/j.dr.2015.12.001

Schwichow, M., Christoph, S., Boone, W. J., & Härtig, H. (2016b). The impact of sub-skills and item content on students' skills with regard to the control-of-variables strategy. *International Journal of Science Education, 38*(2), 216–237. https://doi.org/10.1080/09500693.2015.1137651

Schwichow, M., Osterhaus, C., & Edelsbrunner, P. A. (2020). The relation between the control-of-variables strategy and content knowledge in physics in secondary school. *Contemporary Educational Psychology, 63*, Article 101923. https://doi.org/10.1016/j.cedpsych.2020.101923

Siler, S. A., Klahr, D., Proctor, R. W., & Capaldi, E. J. (2012). Detecting, classifying, and remediating: Children\'s explicit and implicit misconceptions about experimental design. *Psychology of science* (pp. 137–180). Oxford University Press. https://doi.org/10.1093/acprof:OSo/9780199753628.003.0007

Song, J., & Black, P. J. (1992). The effects of concept requirements and task contexts on pupils' performance in control of variables. *International Journal of Science Education, 14*(1), 83–93. https://doi.org/10.1080/0950069920140108

Spruijt, A. M., Ziermans, T. B., Dekker, M. C., & Swaab, H. (2020). Educating parents to enhance children\'s reasoning abilities: A focus on questioning style. *Journal of Applied Developmental Psychology, 66*, Article 101102. https://doi.org/10.1016/j.appdev.2019.101102. March 2019.

Stender, A., Schwichow, M., Zimmerman, C., & Härtig, H. (2018). Making inquiry-based science learning visible: The influence of CVS and cognitive skills on content knowledge learning in guided inquiry. *International Journal of Science Education, 40*(15), 1812–1831. https://doi.org/10.1080/09500693.2018.1504346

Stevenson, C. E., Hickendorff, M., Resing, W. C. M., Heiser, W. J., & de Boeck, P. A. L. (2013). Explanatory item response modeling of children\'s change on a dynamic test of analogical reasoning. *Intelligence, 41*(3), 157–168. https://doi.org/10.1016/j.intell.2013.01.003

Strand-Cary, M., & Klahr, D. (2008). Developing elementary science skills: Instructional effectiveness and path independence. *Cognitive Development, 23*(4), 488–511. https://doi.org/10.1016/j.cogdev.2008.09.005

Taber, K. S. (2018). The use of Cronbach\'s alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Tairab, H. H. (2015). Assessing students' understanding of control of variables across three grade levels and gender. *International Education Studies, 9*(1), 44–54. https://doi.org/10.5539/ies.v9n1p44

Tekkaya, C., & Yenilmez, A. (2006). Relationships among measures of learning orientation, reasoning ability, and conceptual understanding of photosynthesis and respiration in plants for grade 8 males and females. *Journal of Elementary Science Education, 18*(1), 1–14. https://doi.org/10.1007/BF03170650

Thompson, E. D., Bowling, B. V., & Markle, R. E. (2018). Predicting student success in a major\'s introductory biology course via logistic regression analysis of scientific reasoning ability and mathematics scores. *Research in Science Education, 48*(1), 151–163. https://doi.org/10.1007/s11165-016-9563-5

Thuneberg, H., Hautamäki, J., & Hotulainen, R. (2015). Scientific reasoning, school achievement and gender: A multilevel study of between and within school effects in Finland. *Scandinavian Journal of Educational Research, 59*(3), 337–356. https://doi.org/10.1080/00313831.2014.904426

Tytler, R., & Peterson, S. (2003). Tracing young children\'s scientific reasoning. *Research in Science Education* (Vol. 33).

UNESCO. (2011). World Data on Education. http://www.ibe.unesco.org/fileadmin/user_upload/Publications/WDE/2010/pdf-versions/Viet_Nam.pdf. (accessed on 26 March 2021).

Valanides, N. (1997). Cognitive abilities among twelfth-grade students: Implications for science teaching. *Educational Research and Evaluation, 3*(2), 160–186. https://doi.org/10.1080/1380361970030204

Van Vo, D., & Csapó, B. (2020). Development of inductive reasoning in students across school grade levels. *Thinking Skills and Creativity, 37*(2020), Article 100699. https://doi.org/10.1016/j.tsc.2020.100699

Van Vo, D., & Csapó, B. (2021a). Exploring students' science motivation across grade levels and the role of inductive reasoning in science motivation. *European Journal of Psychology of Education*. https://doi.org/10.1007/s10212-021-00568-8

Van Vo, D., & Csapó, B. (2021b). Development of scientific reasoning test measuring control of variables strategy in physics for high school students: Evidence of validity and latent predictors of item difficulty. *International Journal of Science Education*, 1–21. https://doi.org/10.1080/09500693.2021.1957515

van der Graaf, J., Segers, E., & Verhoeven, L. (2015). Scientific reasoning abilities in kindergarten: Dynamic assessment of the control of variables strategy. *Instructional Science, 43*(3), 381–400. https://doi.org/10.1007/s11251-015-9344-y

Van Vo, D., & Csapó, B. (2023). Exploring inductive reasoning, scientific reasoning and science motivation, and their role in predicting STEM achievement across grade levels. *International Journal of Science and Mathematics Education*. https://doi.org/10.1007/s10763-022-10349-4, 0123456789.

Van Vo D. (2022). Assessing inductive reasoning, scientific reasoning and science motivation: Cross-ectional studyies in Vietnamese context. Unpublished doctoral dissertation. University of Szeged. 10.14232/phd.11134.

Vietnam National Assembly. (2006). Luật Giáo dục 2005 [Education Law 2005]. The Publication of Labour and Society.

Vorholzer, A., von Aufschnaiter, C., & Boone, W. J. (2020). Fostering upper secondary students' ability to engage in practices of scientific investigation: A comparative analysis of an explicit and an implicit instructional approach. *Research in Science Education, 50*(1), 333–359. https://doi.org/10.1007/s11165-018-9691-1

Wagensveld, B., Segers, E., Kleemans, T., & Verhoeven, L. (2015). Child predictors of learning to control variables via instruction or self-discovery. *Instructional Science, 43*(3), 365–379. https://doi.org/10.1007/s11251-014-9334-5

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag. https://ggplot2.tidyverse.org.

Wood, K. E., Koenig, K., & Owens, L. (2018). Development of student abilities in control of variables at a two year college. *AURCO Journal, 24*, 164–179.

Wu, M., Adams, R., Wilson, M., & Haldane, S. (2007). *ACER conquest 2.0 manual*. Camberwell: ACER Pres.

Zhou, S., Han, J., Koenig, K., Raplinger, A., & Pi, Y. (2016). Assessment of scientific reasoning : The effects of task context, data, and design on student reasoning in control of variables. *Thinking Skills and Creativity, 19*, 175–187. https://doi.org/10.1016/j.tsc.2015.11.004

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27*(2), 172–223. https://doi.org/10.1016/j.dr.2006.12.001

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*(1), 1–28. https://doi.org/10.1111/j.1745-3984.1999.tb00543.x