UNIVERSITY OF SZEGED
DOCTORAL SCHOOL OF EDUCATION

DE VAN VO

ASSESSING INDUCTIVE REASONING, SCIENTIFIC REASONING AND SCIENCE MOTIVATION: CROSS-SECTIONAL STUDIES IN VIETNAMESE CONTEXT

DOCTORAL DISSERTATION

SUPERVISOR:
Prof. Dr. BENŐ CSAPÓ

SZEGED, HUNGARY, 2022

ABSTRACT

The present research aims to explore developmental patterns in reasoning abilities and science motivation as well as their relationship with students' background variables in different grade levels. Three cross-sectional investigations assessed 2241 students in 16 public schools in Vietnam.

The first study explored development in students' inductive reasoning and science motivation in 5th, 7th, 9th, and 11th graders. The findings showed that the older age groups tended to achieve higher scores than their younger counterparts on the inductive reasoning test, but students' science motivation gradually dropped grade by grade. Multi-model Bayesian inference discovered that inductive reasoning, science performance, and parental involvement were the main predictors of science motivation. Furthermore, path analyses showed that inductive reasoning has an indirect effect on science motivation via a science performance variable. The second study, which assessed reasoning abilities and science motivation of 6th, 8th, 10th, and 11th graders, revealed that students' scores on inductive reasoning and scientific reasoning tests improved across grade levels. Path analyses suggested that inductive reasoning, scientific reasoning, father's education, and parental involvement are chief factors in predicting STEM achievement. The third study developed a scientific reasoning test to measure control of variables strategy in physics in the 8th to 12th-grade students. The reliability and validity of the test were confirmed by the Rasch model. Development of control of variables strategy in physics in students was observed, but students' motivation toward learning physics reduced slightly through grade levels.

No significant difference was found between males and females in reasoning abilities and science motivation, but a significant difference was indicated in physics motivation, favouring boys. Effects of multimedia on psychometric characteristics of the cognitive tests were confirmed by measurement invariance regarding administration modes, supporting the online groups on the item bundles constructed of figure-related materials. The implications for enhancing reasoning proficiencies and science motivation are also discussed further in each study.

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my supervisor, Prof. Dr. Benő Csapó, who has taken me into this exciting topic, provided me with thorough assistance and motivation to accomplish my PhD journey and for all his support from the early phases of scholarship application. I would like to thank the Stipendium Hungaricum scholarship for their financial support.

I am grateful to Prof. Tibor Vidákovich for his course of educational assessment, Prof. Gyöngyvér Molnár for her useful lessons in Rasch model and technology-based assessment, Prof. Józsa Krisztián for his excellent course in research methods, all professors and staff at the Institute of Education in University of Szeged for their support. I also would like to thank my classmates and the PhD students in the Doctoral School of Education for all their supportive discussions. To my beloved friends in Szeged, I would like to thank you all for the happy times we spent together that I have never forgotten in my life.

I would like to say big thanks to my family members for their unconditional sacrifices and their encouragement during my steps. My sincere thanks also go to my colleagues at An Giang University - Vietnam National University Ho Chi Minh City. I would like to thank the principals and teachers in the participating schools for their cooperation.

I am thankful to the Center for Research on Learning and Instruction, University of Szeged for the opportunity of using the eDia platform, and also to the Klebelsberg Library in the University of Szeged for academic support during my PhD journey. Most importantly, I would like to thank the students for their voluntary participation in my research.

TABLE OF CONTENTS

ABBREVIATIONS

AAAS	American Association for the Advancement of Science
AG	Achievement goals
AGE	Student age
AL	Active learning strategies
ANOVA	Analysis of variance
BMA	Bayesian model averaging
CFA	Confirmatory factor analysis
CFI	Comparative fit index
CL	Students’ confidence in learning physics
CVS	Control of variables strategy
CVSP	Control of variables strategy in physics
DBF	Differential bundle functioning
DE	Deductive reasoning
DIF	Differential item functioning
eDia	Electronic Diagnostic Assessment
ET	Students’ views on engaging in physics lessons
FA	Figure analogies
FE	Father’s education level
FS	Figure series completion
GPA	Grade point average
GR	Grade level
ICT	Information and communication technology
ID	Identifying controlled experiments
IN	Interpreting the outcome of a controlled experiment
IR	Inductive reasoning
LE	Learning environment stimulation
LL	Student like learning physics
LV	Science learning value
M	Mean
ME	Mother’s education level
MLE	Maximum likelihood estimation

MOET	Ministry of Education and Training
N	Number
NA	Number analogies
NS	Number series completion
OECD	Organization for Economic Co-operation and Development
PCA	Principal component analysis
PH	Physics test in the previous semester
PI	Parental involvement in schooling
PISA	The Programme for International Student Assessment
PP	Paper-and-pencil
PRISMA	Preferred Items for Systematic Reviews and Meta-Analysis
RMSEA	Root mean square error of approximation
RPM	Raven's Standard Progressive Matrices
SD	Standard deviation
SE	Self-efficacy
SEM	Structural equation modelling
SIBTEST	Simultaneous item bias test
SM	Science motivation
SMTSL	Students' motivation toward science learning
SR	Scientific reasoning
TBA	Technology-based assessment
TIMSS	Trends in International Mathematics and Science Study
TLI	Tucker-Lewis index
Tukey's HSD	Tukey's honestly significant difference
UN	Understanding the determinacy of confounded experiments
WRMR	Weighted root mean square residual

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

1.1 Introduction

In light of contemporary frameworks for the 21st century, researchers, employers and policy makers called attention to the need for competencies in communication, collaboration, digital-related competencies and social skills. Creativity and problem-solving have been considered as key competencies in the 21st century (Binkley et al., 2012; Chu et al., 2017; Voogt & Roblin, 2012). Human reasoning is a considerable topic of widespread study since the time of Aristotle and still a much interesting field in psychological, theoretical, and empirical research today. Reasoning plays an essential role in both educational contexts and the workplace in modern society, where more information and knowledge created in a shorter time leads us to an increasing pressure to manage. Amongst forms of reasoning, inductive reasoning (IR), one of the main elements of fluid intelligence, and scientific reasoning (SR) are increasingly considered in school contexts.

IR has been one of the most frequently studied constructs in reasoning field, and recent social changes and expectations concerning the outcomes of education have further highlighted its relevance. Previous studies have suggested that there is a close association between IR, on the one hand, and problem-solving and academic success, on the other (Csapó, 1997; Molnár et al., 2013; Schweizer et al., 2013). IR has been shown to play a central role in a wide range of learning activities (Hamers et al., 1998). As a main component of intelligence, IR assists students in decision-making and establishing causal relationships (Leighton & Sternberg, 2004).

Meanwhile, the term SR has been used as a domain-specific approach of reasoning skills in science subjects. SR refers to both in the science subject and in separated STEM subjects (Zimmerman, 2000). According to Lawson (2009), content knowledge and SR are two foundational pillars of scientific literacy, so science education mainly aims to develop both content knowledge and SR ability of students.

Various studies have also provided evidence of the direct and indirect relationships between motivation towards science learning and science achievement among students from primary to tertiary education (e.g. Chan & Norlizah, 2018; Dermitzaki et al., 2013; Glynn et al., 2009, 2011; Kambeyo, 2018). Academic motivation has been considered as a supportive factor in enhancing students' performance. Motivation in learning science is not only the main factor in predicting science attitude, but it can also predict academic success (Cavas, 2011;

Dermitzaki et al., 2013). Science motivation has a consequential influence on reasoning ability (Kambeyo, 2018; Tee et al., 2018) and learning strategies (Patrick et al., 2009), and it plays as a predictive factor of science performance (Chan & Norlizah, 2018; Patrick et al., 2009). Research on relationship between IR and SM and interaction of SR with other factors in predicting student motivation and performance in schools are still scarce in educational context. Additionally, a broad range of measurement data, through longitudinal studies (e.g. Hwang et al., 2016; Robinson et al., 2019) and cross-sectional studies (e.g. Csapó, 1997; Dorfman & Fortus, 2019; Józsa et al., 2017), often offer a deeper understanding of individuals' differences in developmental pattern in students' cognitive abilities and motivation, which plays an important role in personalized support in enhancing children's academic performance in educational contexts.

In Vietnam, thinking skills are embedded in the core curricula at the general educational level. The national educational program aims to develop students' thinking proficiency through implicit discipline curricula rather than designing specific programs explicitly. Consequently, there are no explicit programs that focus on developing thinking skills. Though one core criterion for completing general education is for children to pass the national high school examination, little or no attention has been paid on investigating students' thinking skills on this examination. However, Vietnamese students achieved high scores in the Program for International Student Assessment (PISA) by the Organization for Economic Co-operation and Development (OECD), which mainly measures content-related knowledge, skills, and their application in practical contexts. For example, in the 2015 PISA cycle, Vietnamese students achieved significantly better than the OECD average in science, with a mean score of 525 (OECD mean: 493), and not significantly different from the OECD average in science and reading (OECD, 2016). Vietnam is also among the countries where motivation for learning science is above the OECD average (OECD, 2016), but PISA surveys only show the results of 15-year-old students' motivation. Further studies are therefore needed to gain insights into patterns of students' motivation toward learning science in other grade levels. Vietnam did not participate in the PISA problem-solving assessments, and relatively little is known about how schools have succeeded in developing thinking in general. Moreover, students learn science as a separate subject starting from their first year of secondary education (6th grade), and one of the requirements of the national education programs is to ensure that around 10% of the physics curriculum in high schools includes experimental lab activities (MOET, 2009). Nonetheless, students' scientific reasoning has been rarely assessed on tests either at the regional levels or at the national levels (e.g., national examinations) in Vietnam.

No empirical studies were broadly conducted to explore general reasoning skills and SM in Vietnam. With this as a backdrop, the main goals of our investigations are to validate the adapted test instruments for assessing reasoning skills and science motivation and to explore the patterns of students' reasoning and SM across grade levels. Gender differences in reasoning and SM are observed in different grade cohorts. We also examine interactions of relevant factors (e.g., age, gender, science achievement, parental involvement in schoolwork and parents' education level) in predicting reasoning capacities, SM and STEM achievement in Vietnamese students. The present study is expected to draw a partial picture of students' thinking skills and motivation toward science learning and to provide initial empirical foundations for further research as well as to support the developmental work of teachers.

1.2 The context of study

The national education system in Vietnam consists of four levels (Vietnam National Assembly, 2006). Educational levels and training degrees in the national education system are as follows: early childhood education with nursery and kindergarten; general education with primary education, lower secondary education, and upper secondary education; professional education with professional secondary education and vocational training; and higher education with college undergraduate, master's and doctoral courses. Within general education, primary education is compulsory for all children aged 6 to 10 years. Lower secondary education lasts four years (6th to 9th grades) for children aged 11 to 15 years, while students aged 15 to 18 years can enrol in upper secondary education (Figure 1.1).

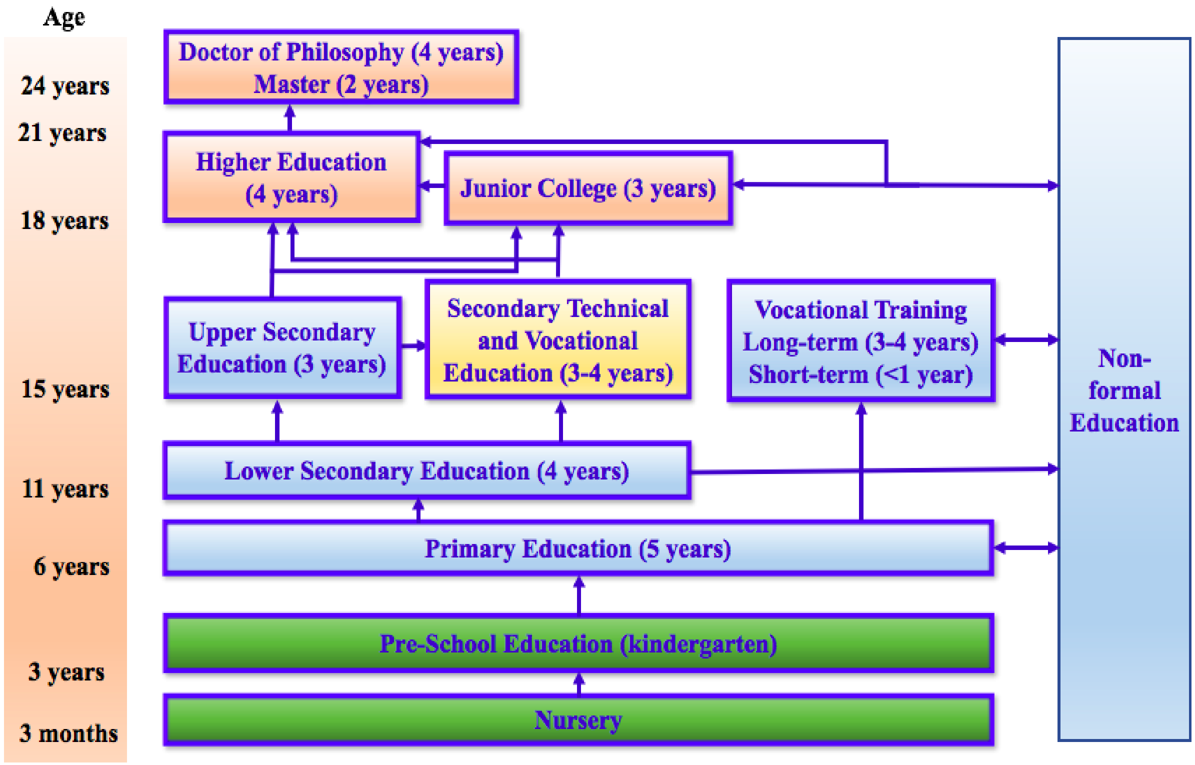

Age
24 years
21 years
18 years
15 years
11 years
6 years
3 years
3 months
Doctor of Philosophy (4 years)
Master (2 years)
Higher Education (4 years)
Junior College (3 years)
Upper Secondary Education (3 years)
Secondary Technical and Vocational Education (3-4 years)
Vocational Training
Long-term (3-4 years)
Short-term (<1 year)
Lower Secondary Education (4 years)
Primary Education (5 years)
Pre-School Education (kindergarten)
Nursery
Non-formal Education

Figure 1.1. The structure of the national educational system in Vietnam.

The core subjects in primary schools include Vietnamese language, Mathematics, Natural and Social Sciences, Moral Education (civics), Physical Education, Arts and Foreign language. In secondary education, the compulsory subjects include Vietnamese language and Literature, History, Geography, Civics, Art, Music, Physical Education, Foreign language, Mathematics, Physics, Chemistry, Biology and Technology. Vocational and social activities are also scheduled in in-class or out-door settings during the academic year. To continue upper secondary education, 9th-grade students have to be successful in a provincial examination. After completing high school, 12th graders are required to take the National High School Graduation Examination to earn a diploma, called the High School Graduation Certificate before enrolling to a university or college.

Schools in Vietnam have focused on forming three major general competencies: autonomous and independent learning, communication and cooperation, and creativity and problem-solving (MOET, 2017). Although teachers are fully aware of the benefits of teaching to develop thinking skills, they regard it as extraneous to the requirements of testing, the criteria for teacher evaluation and the general expectations of many parents (Du, 2015). Consequently, teaching and learning in Vietnamese schools have been criticized for being exam-based. Though such focus of teaching and learning did help students pass the exams with high results, their practical abilities of applying their acquired knowledge seemed to be limited (Nhat et al., 2018).

The current general education curriculum has been introduced throughout the country starting from the 2002–2003 school year. Objectives, content, curricula, textbooks and regulations on completion requirements as well as other relevant matters are implemented similarly in both non-public and public schools (UNESCO, 2011). Therefore, the cross-sectional assessments can provide useful information about how well learning science and teaching thinking skills can be integrated into subject-specific areas at different grade levels. This may be meaningful in practice vis-à-vis boosting students' reasoning capacities and motivation and in proposing improved programmes in future.

1.3 The present dissertation

Our studies aim to address some aspects of assessment of reasoning capacity and science motivation of students. First, we review and synthesize the findings from previous empirical studies related to reasoning capacities and science motivation. Such research is expected to outline the main trends and identify the various measures that are widely used when assessing reasoning skills around the world. The literature review attempts to examine the relationships between theoretical concepts and the empirical paradigms regarding the construction of the instrument tests, psychometric approaches and technology issues to measure reasoning capacity in school settings. The relationship between cognitive abilities and science motivation was reviewed from the former studies. Consequently, this may provide insights into the current trend of assessing reasoning and motivation, the main relevant factors (demography and cognition) influencing the reasoning ability and their interaction in explaining academic achievement.

To this end, a set of three cross-sectional studies were proposed to assess reasoning abilities and science motivation of secondary school students. The first empirical study focuses on exploring developmental patterns of IR and SM across grade levels and their association in predicting children's science achievement. Examining relationships of IR and other latent factors of SM offers an in-depth information into reasoning and motivation among individuals, which plays an important role in personalized learning support in enhancing children's academic success in schools.

The second cross-sectional investigation focuses on exploring the developmental curves of inductive reasoning, scientific reasoning abilities, and their roles in predicting students' performance in learning STEM disciplines across grade levels. Students with a better IR ability are likely to earn higher scores on SR tests (Csapó, 1999) and be more active in inquiry activities in learning science (Gerber et al., 2001), and vice versa, IR is a foundational skill in boosting SR (Adey et al., 2007). Assessment of the IR and SR and their interaction with background variables may offer a deeper insight into the relationship cognitive capacities and motivation. It also partly provides evidence for evaluating the extent of the present curricula in terms of optimum cognitive development in school-age children.

In line with the first two studies, in the third research, we considered examining students' SR and motivation in learning physics. Control of variables strategy (CVS) is a leading component of SR related to domain-general experimentation to evaluate an experimental system to draw valid conclusions. The purposes of this study are to develop, validate the test to measure CVS in physics, and explore development of this ability in secondary school students in Vietnamese context. The study also outlines basic rationales of assessment of CVS and explores latent predictors of item difficulty as a practical reference for teachers who consider designing assessment-based learning activities in classroom and hands-on tasks in school labs.

Additionally, technology-based assessment (TBA) has become a contemporary trend in education and has gradually replaced paper-and-pencil (PP) assessment in recent years. Across these empirical studies, the effects of administration modes are also observed by testing measurement equivalence in the test, task and item levels. This may show the initial correspondence of feasibility of applying TBA in Vietnamese context.

1.4 The construct of the dissertation

As mentioned in the main research purposes above, the thesis, which mainly focuses on assessing the aspects of IR, SR and SM, is organized into eight chapters.

Chapter 1 highlights general core facets of research problems in the project. It provides main points from former studies in reasoning and science motivation in educational settings around the world as well as introduces the study context in Vietnam. The first chapter shows a general sketch of our research project.

Chapter 2 presents the main characteristics and trends in measuring IR and SR around the world through a systematic review and thematic analysis with 63 empirical studies from 1997 to 2020. The non-verbal analogical problem was indicated as the most popular task in evaluating IR, while control of variables task was most frequently applied in assessing SR in school contexts. These findings also show that students' reasoning abilities can be efficiently enhanced through school disciplines. In addition, motivational factors and assessment of science motivation were underlined in this chapter. Although the underlying concepts of motivation are controversial, most of the studies reported that students' motivation toward science learning had a close relationship with their academic performance in schools. A part of these findings has been accepted to publish in *International Journal of Innovation and Learning* (Van Vo & Csapó, *in press*)

In Chapter 3, we discuss the main methods applied in three empirical studies. The chapter introduces a cross-sectional investigation as a main approach in a series of these studies. The test instruments are presented in this chapter, including the IR test, SR test, CVS test in physics, science motivation questionnaire, student motivation and attitude toward physics learning and the background questionnaire. Moreover, we summarized the main methods for data analysis and interpretation such as the principal component analysis (PCA), confirmatory factor analysis (CFA), Rasch model measurement, differential item difficulty (DIF), path analysis and multiple linear regression models with Bayesian model averaging (BMA) approach.

Chapter 4 illustrates the first empirical study that aims to investigate the development of IR and patterns of students' motivation across the 5th, 7th, 9th, and 11th grades. The initial findings showed that students' performance improved steadily grade by grade, whereas their motivation toward learning science gradually fell through the grade levels. The main factors contributing to individual IR capacity and science motivation are also discussed in this chapter. The findings of development of IR across grade levels had published in the *Thinking Skills and Creativity* journal (Van Vo & Csapó, 2020), while the results related to an investigation of students' science motivation and its interaction with IR, were accepted as a research article in the *European Journal of Psychology of Education* (Van Vo & Csapó, 2021a).

The findings of the second study are reported in Chapter 5, which showed the developmental curves of IR, SR capacities and SM in students in the 6th, 8th, 10th, and 11th grades. The developmental curves in IR and SR were simulated across grade cohorts, while the students' SM seems to reduce in the upper grade cohorts. Furthermore, the predictive roles of IR, SR and SM on children's STEM achievement are discussed in this chapter.

Chapter 6 details the third empirical study which considers assessing scientific reasoning in CVS in physics in secondary school students. This chapter presents the development and validity of the CVS test to measure three subskills of CVS related to content knowledge in basic physics (mechanics, thermodynamics, and electricity). A part of these results has been published in the *International Journal of Science Education* (Van Vo & Csapó, 2021b). In general, students' CVS proficiency progressed during secondary education years, but their motivation toward physics learning tended to reduce during secondary education level.

Relationships between CVS, motivation and content knowledge in physics were also examined in this study.

Since tudents took the cognitive tests in either PP or online formats, Chapter 7 discusses effects of media delivery modes in terms of students' performance and tests' psychometric properties across three empirical data. The examinations of the equivalence of dual test versions were conducted in both internal structural factors and DIF analyses. We employed classical statistics and Rasch model measurement to compare students' score results and psychometric properties of the tests at the test, task, and item levels.

Finally, Chapter 8 summarizes the general conclusion and discussion across these investigations. The limitations and future directions are also discussed in the last chapter.

CHAPTER 2. REVIEW OF THE LITERATURE

2.1 Assessing inductive reasoning and scientific reasoning

2.1.1 Introduction

Globalization and the rapid development of Information and Communications Technology (ICT) are continuously transforming in most aspects of modern life, work, and school. New conceptions of knowledge and students' skills are expected to meet the demands of future jobs. In a comparison of a number of frameworks for 21st-century competencies, Voogt and Roblin (2012) strongly recommended the need for competencies, such as communication, collaboration, ICT-related competencies and social awareness. Creativity, critical thinking, and problem-solving have also become important competencies in the 21st century.

Assessment plays a pivotal role in the teaching and learning process as well as in school administration. Assessment involves the process of measuring, collecting, and using evidence about the outcomes of students' learning, provides feedback for students in their learning processes, and helps teachers and other stakeholders make evidence-based decisions (Hattie & Timperley, 2007). Traditional assessment methods seem to fail to measure high-level skills, attitudes, and even knowledge of collaborative learning in a fast-changing world (Griffin et al., 2012). Most of the researchers have agreed that current assessment models, which are mostly focused on measuring separate knowledge, fail to assess 21st-century competencies, and called for new assessment established for an authentic and complex task (Pepper, 2011; Voogt & Roblin, 2012). Learners' abilities are implicitly present in a range of variables and a combination of several factors. The structure of the new assessment should involve dynamic procedures and rich-environment instruments to collect several relevant types of information about the learners' application of competencies, which consist of outcomes, processes, and rationales (Pepper, 2011). TBA will become a major feature in the educational setting. This coming technological transformation makes it possible to assess wider-ranging student performance, including assessing core reasoning skills (Osborne, 2013).

Research on assessing reasoning has been an established field of inquiry for many decades, and improved instruments have made it possible to measure a variety of reasoning skills and explore their development and the links between relevant factors. Empirical research has identified the importance of reasoning skills in the operation of higher-order cognitive skills (Csapó, 1997; Molnár et al., 2013; Schweizer et al., 2013) and in learning most school subjects, such as mathematics (Nunes & Csapó, 2011), science (Adey, & Csapó, 2012; Hamers et al., 1998) and foreign languages (Nikolov & Csapó, 2018; Soodmand Afshar et al, 2014) as well as other subjects. Reasoning also assists students in decision-making, problem-solving, understanding causal relationships among different domains, and achieving academic success in school curricula. Some previous studies have focused on reviewing the development of SR and problem-solving strategies (Zimmerman, 2000) while others investigated the role of IR enhancement (Klauer and Phye, 2008), scientific reasoning (e.g. Engelmann et al., 2016) and control of variables strategy (Schwichow et al., 2016) through reporting findings of the intervention studies. Still, others have analyzed instruments for assessing SR (Opitz et al., 2017). Though several theoretical frameworks and comprehensive studies have explored the place of reasoning skills in general intelligence (Carroll, 1993; Leighton & Sternberg, 2004), a systematic review from empirical studies has not been thoroughly conducted in educational contexts.

Thus, the section synthesizes findings of a systematic literature review to outline the main trends and identify the various measures that are used to assess reasoning in educating sittings around the world. The review also focuses on exploring the relationship between reasoning and academic performance and main relevant factors under different cultural conditions. Consequently, we hope to provide insights into the current trends in assessing reasoning and recommend possibilities for future studies in educational settings. The following five research questions guide our review:

(1) How is reasoning assessed in school contexts around the world?

(2) What are the most popular problem tasks used to assess reasoning in educational contexts?

(3) What are the trends of applying the psychometric modelling and technology in assessing reasoning?

(4) To what extent are learners’ reasoning abilities influenced by gender, age differences and educational programs?

(5) To what extent do reasoning and academic performance relate to other relevant factors?

2.1.2 The concepts of reasoning

Human reasoning has been a focal topic of widespread study since the time of Aristotle. It continues to be the main concern of both psychological theory and empirical research nowadays. In the literature, reasoning has undergone a variety of definitions. Evans (1993) defined reasoning as “the central activity in intelligent thinking. It is the process by which knowledge is applied to achieve most of our goals” (p. 561). Perret (2015) highlighted that reasoning does not operate automatically; this process requires utilizing attentional resources and intellectual effort to justify inferences. Additionally, Sternberg (1986) viewed reasoning as an extensive mental phenomenon which can be measured through problems besides other skills. In reasoning, we are from "what is already known to infer a new conclusion or to evaluate a proposed conclusion" (Sternberg & Sternberg, 2012, p. 507), and it is considered an essential component of critical thinking (Bonney & Sternberg, 2011). Although many different definitions of reasoning can be found in the literature, it can broadly be defined as the goal-driven process of drawing conclusions which inform problem-solving and decision-making efforts (Leighton & Sternberg, 2004).

One of the most popular criteria to classify forms of reasoning distinguishes two main types: deductive and inductive reasoning (Sternberg, 1986; Sternberg & Sternberg, 2012; Wesiak, 2003). IR, one of the most widely researched cognitive skills, is a cognitive process based on particular facts or individual cases to induce a general conclusion (Adey & Csapó, 2012; Sternberg & Sternberg, 2012). It plays a significant role in understanding science and applying knowledge in unfamiliar situations (Csapó, 1997). It has been regarded as one of the seven primary mental abilities to contribute to intelligent behaviour (Kinshuk et al., 2006). In many cases, it provided a valid foundation for the understanding of regularities. As mentioned by Klauer and Phye (2008), new concepts and knowledge in our daily life are generated based on regularities and uniformities. Therefore, IR plays an more important role when complex or unfamiliar problems occur, where no specific content knowledge might be applicable. Inductive processes can be applied to generate hypothetical rules based on observation, and these relational systems of problems can be modelled (Perret, 2015). In contrast to IR, deductive reasoning is a cognitive process that concludes a logically specific conclusion from general information (Sternberg & Sternberg, 2012). The development of deductive reasoning was closely linked to awareness of cognitive processes and cognitive control (Demetriou et al., 2011).

With a growing emphasis on science education, a further form of reasoning has become a focal research topic, and the term SR has been used in a domain-specific approach in science subjects. SR is defined as an active procedure of interrelating a series of cognitive and metacognitive processes (i.e. reasoning) to generate, test, and revise theories and hypotheses (Zimmerman, 2007). As broadly defined by Osborne (2013), scientific reasoning is the use of reasoning in science, so it involves reasoning with requisite scientific knowledge Inquiry activities support enhancing the process of knowledge acquisition and knowledge application, which contributes to the development of SR abilities (Lawson, 2009). SR plays a core role in science subjects in general and in separate subjects (mathematics, physics, biology and chemistry) (Zimmerman, 2000). Although reasoning abilities are naturally present in early childhood, these capacities can be cultivated in the long term through the stimuli and information process in educational contexts (Köksal-tuncer & Sodian, 2018; Zimmerman, 2000, 2007). Therefore, the main goals of an interdisciplinary approach among subjects tend to focus on developing both content knowledge and SR ability in school-age children.

2.1.3 Assessing reasoning abilities

Assessing reasoning often involves a series of scenarios in which students are observed in a step-by-step process through their responses in each situation. In the past, an individual interview method was regularly applied to examine reasoning or thinking ability via Piaget's clinical technique (Adey & Csapó, 2012). Researchers then developed reliable and valid classroom tests of cognitive development in PP and computer-based formats. A variety of tasks have been developed to assess the kinds of reasoning, considering the latent factors contributing to individual differences in reasoning. Generally, most of the test instruments have been developed and adapted based on certain popular tasks as follows.

Regarding inductive reasoning, the problem tasks may be grouped into four main groups: analogies, series completions, classifications and geometric matrices (verbal, numeral, and figural element) (Adey & Csapó, 2012; Klauer & Phye, 2008; Sternberg, 1986; Sternberg & Sternberg, 2012). Analogy tasks take the form "A is to B as C is to D" and are most commonly used in intelligence tests (Sternberg, 1986). A standard format has a structure of a three-term stem ("A is to B as C is?") and a set of answer options. To get the correct alternative, test-takers should find the relation between the terms A and B, then apply the rules to term C, which has the same rules to the former pair stems. The item difficulty of the analogy tasks may depend on the type of semantic relation between the pairs, the number of relevant elements or concepts in the stem, the detectability of the given relation and even the materials of elements. A series completion task requires participants to extrapolate to find a missing member of a given series of elements (numbers, letters, words, or geometric figures). In the task, a set of elements is ordered according to one or more relations between the elements. The person being tested must explore the rules from the relationships in the series. The number of elements, pattern presentation, and number of relation rules can influence the item difficulty of series completion problems (Wesiak, 2003). Classification tasks take several forms. The task is composed of a series of elements in which one element is the correct alternative. The basis for solving classification items needs similar approach as in analogy and series completion problems: test takers should explore the comprehension of meanings and of semantic relations among given elements. Participants discover the rules in a set of elements and may be requested to identify elements (words, figures, or numbers) that belong or do not belong to the rules. Finally, geometric-figural matrix tasks refer to a series of figures which are arranged in a matrix. These tasks are usually composed of a diagram with given figures set into rows and columns with one cell missing. Participants need to extrapolate the relationships among the figures in the matrix (rows and columns) to find the figures in the missing cell which correctly fits the whole matrix. Geometric-figural matrix items are most frequently found in intelligence tests such as Raven's Standard Progressive Matrices (RPM) (McCallum, 2017).

Deductive reasoning (DR) tasks can be grouped into two types, conditional and syllogistic reasoning (Sternberg & Sternberg, 2012). In conditional reasoning, a problem-solver completes the task using an if-then proposition to draw a valid conclusion. Wason's four-card task is an example of conditional reasoning. Syllogisms are deductive arguments concluded from two premises. The most well-known syllogism is the categorical syllogism.

With regard to scientific reasoning, several kinds of problem tasks have been proposed, most frequently named conservation, control of variables, proportions and ratios, probability, correlational reasoning and hypothetical-deductive reasoning (e.g. Adey & Csapó, 2012; Han, 2013; Lawson, 2000; Lawson et al., 2000). Conservation requires the ability to identify a quantity that remains unchanged if nothing is added or taken away although the appearance of an object is changed. Conservation may be in number, matter (mass), weight, and volume of liquid (Adey & Csapó, 2012). The control of variables task consists of a complex reasoning pattern or strategy within several reasoning schemes (Adey & Csapó, 2012). This task is used to determine whether an experiment can reach a conclusive result when some conditions of variables are changed. Proportions and ratios tasks contain some complex analogical and inductive forms of reasoning (Csapó, 1997; Nunes & Csapó, 2011). The task requires a sense of covariation and multiple comparisons in processing several pieces of information. The mathematical relation among quantities may be assumed to be linear or nonlinear. A probability task involves probabilistic inferences based on existing events and estimating likelihoods of future events (Adey & Csapó, 2012). Meanwhile, correlational reasoning tasks require exploring probabilistic relationships between two features or variables appearing in a certain number of cases. The strength of the association may be different based on the ratio of the particular cases. For example, two variables may be either independent of or related to each other. A correlation task is often employed to describe the degree of dependence between two or more variables (Han, 2013). A hypothetical-deductive reasoning task usually starts with a general concept of all latent factors that might affect an outcome and production of a hypothesis. Based on the hypothesis, deductions are reasoned to predict what might occur in an experiment (Han, 2013). This capacity is very important in learning STEM disciplines (Kalinowski & Willoughby, 2019). Additionally, combinatorial reasoning is considered as a type of scientific reasoning task (Adey & Csapó, 2012). It operates by combining and evaluating the possibilities that satisfy the conditions explicitly given or inferred from particular situations. Combinatorial reasoning related to hypothetical-deductive reasoning and probability, applied in in different fields such chemistry, biology, physics, communications, and number theory. As discussed above, the term SR refers applying cognitive abilities, including forms of reasoning in specific-domain of science subjects, so some researchers (e.g., Korom et al., 2017) considered IR tasks which the elements are constructed of science content as a kind of scientific reasoning problem tasks. To solve these tasks, participants are required to execute the IR skill into specific content knowledge of science.

2.1.4 Method of the literature review

We refer to the Preferred Items for Systematic Reviews and Meta-Analysis (PRISMA) statement (Moher et al., 2009) to ensure that our review was systematic, following five steps as summarized in Figure 2.1.

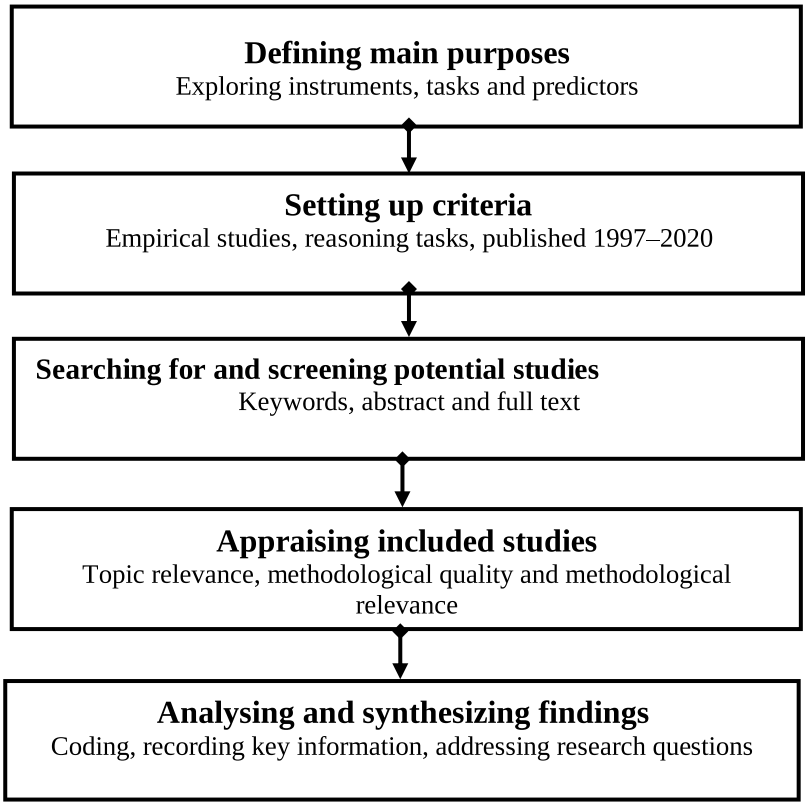

Defining main purposes
Exploring instruments, tasks and predictors
Setting up criteria
Empirical studies, reasoning tasks, published 1997–2020
Searching for and screening potential studies
Keywords, abstract and full text
Appraising included studies
Topic relevance, methodological quality and methodological relevance
Analysing and synthesizing findings
Coding, recording key information, addressing research questions

Figure 2.1. Flowchart for the literature review procedure based on the PRISMA protocol.

We conducted a review of existing studies, focusing on empirical research. The basic principle of selecting an article to be reviewed was based on whether it assesses reasoning or uses reasoning tasks to explore the relation between reasoning and other factors of academic achievement. To draw a general picture of the research topic, we followed the procedure of selecting studies in accordance with the PRISMA statement (Moher et al., 2009). We selected articles or academic documents based on the four inclusion criteria. First, empirical studies assessing the reasoning of students (qualitative, quantitative, or mixed) were included. Second, studies had to apply at least one assessment instrument of reasoning tasks (IR, DR, and SR). Third, they must focus on the relationship between reasoning and academic performance and/or other factors. Fourth, selected papers had to be published in English between 1997 and 2020.

Table 2.1. The results of the initial search in the major electronic databases.

Database	Search Limiters	Hits
JSTOR	Content type: peer-reviewed journals Subject: education Full text: available	425
ERIC	Peer-reviewed journals Publication type: journal articles Full text: available	128
Google Scholar	Perform search with all the words in the title	35
ProQuest	Peer-reviewed journals Dissertations for higher degrees (PhD, EdD) Full text: available	239
ScienceDirect	Article type: research articles Peer-reviewed journals Full text: available	380
	Total	1207

Note. All searches were limited to publications in English since 1997.

Academic papers were searched in electronic databases using the following search terms for the title, keywords, and abstract section: (“reasoning” OR “scientific reasoning” OR “inductive reasoning” OR “deductive reasoning” OR “high order thinking”) AND (“Assessment” OR “technology-based assessment” OR “computer-based assessment” OR “mobile-based assessment”). The major electronic databases were searched in the review: ScienceDirect, Google Scholar, Taylor and Francis, ProQuest, Education Resources Information Center (ERIC) and JSTOR. Each search was limited to empirical studies in peer-reviewed papers. Through the search strategy, around 1207 references were retrieved which are presented in Figure 2.2. After removing over 357 duplicates, the titles and abstracts of the remaining 850 articles were screened individually for pre-selection purposes using the inclusion criteria. Through the screening procedure, 105 empirical studies were selected for further reconsideration based on the four inclusion criteria.

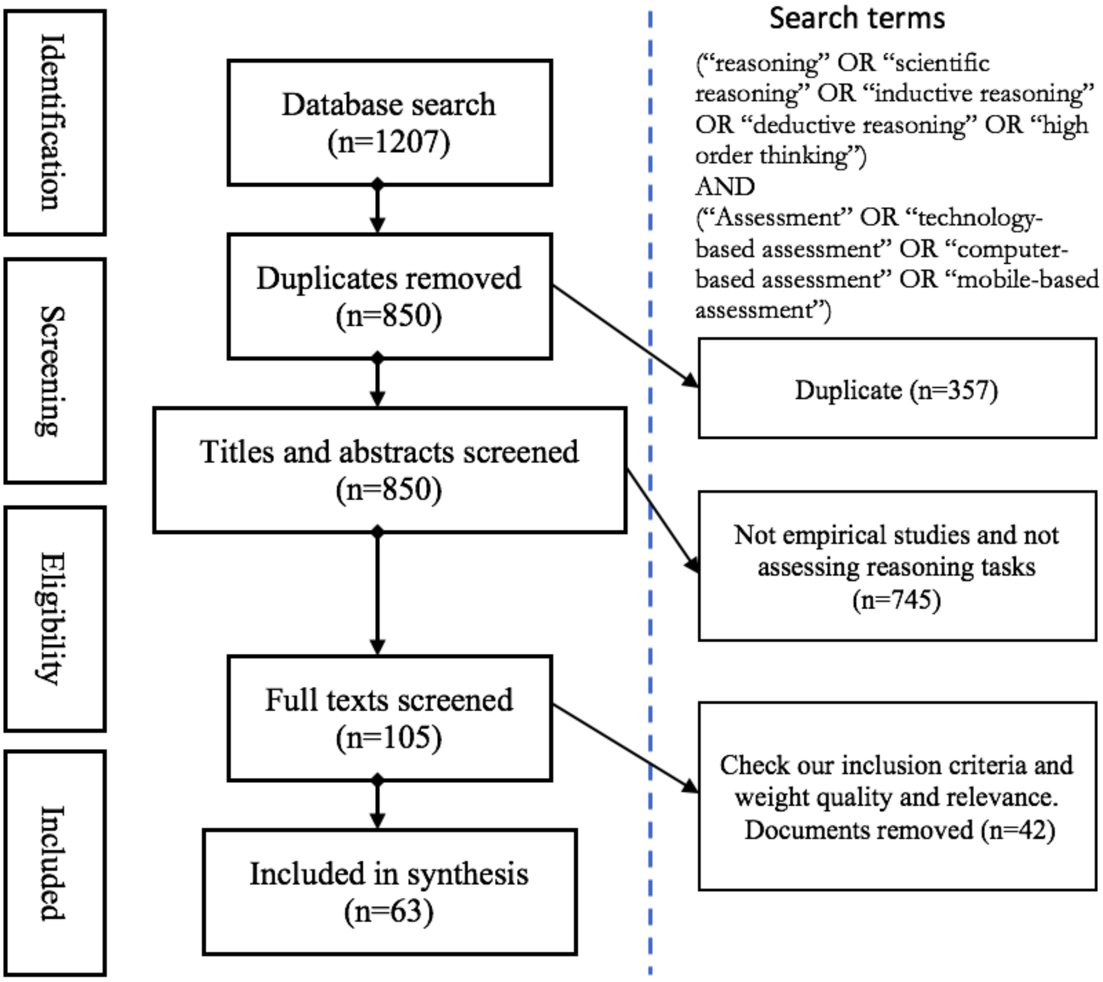

Identification
Screening
Eligibility
Included
Database search (n=1207)
Duplicates removed (n=850)
Titles and abstracts screened (n=850)
Full texts screened (n=105)
Included in synthesis (n=63)
Search terms
("reasoning" OR "scientific reasoning" OR "inductive reasoning" OR "deductive reasoning" OR "high order thinking")
AND
("Assessment" OR "technology-based assessment" OR "computer-based assessment" OR "mobile-based assessment")
Duplicate (n=357)
Not empirical studies and not assessing reasoning tasks (n=745)
Check our inclusion criteria and weight quality and relevance. Documents removed (n=42)

Figure 2.2. Literature search flowchart.

After evaluating pre-selected studies based on the inclusion criteria, we referenced the Weight of Evidence framework (Gough, 2007) to evaluate the remaining 105 studies for their quality and relevance. Following this framework, studies were appraised using three key areas: topic relevance, methodological quality, and methodological relevance. After judging the proposed documents that met the initial inclusion criteria, twelve studies were excluded since they did not meet the quality and relevance standards for the review. Finally, we included 63 empirical studies in the present investigation.

In line with our research purposes, we built a set of summarized features as a template which recorded key information about the sample:

1. year of publication

2. education levels of the participants (we grouped education level into four groups):

a) 1st grade - 5th grade

b) 6th - 9th grades

c) 10th - 12th grades

d) higher education

3. country context

4. operation modes of the assessment (online, paper-and-pencil, or mixed)

5. research design (quantitative, qualitative, or mixed)

6. learning domain and instrument

7. research aims and main findings.

The main themes that appeared across the reviewed studies were grouped. They were later used as the major objectives when designing our studies, particularly when addressing the research questions. A brief overview of the surface features of the selected studies is provided in Appendix A.

2.1.5 The main properties of the studies selected

Figure 2.3 demonstrates the main characteristics of the studies selected, which were published between 1997 and 2020. The first nine years (1997 - 2005) produced only seven papers, then ten articles were published in the following 5-year periods (2006 - 2010). The number of studies doubled in the subsequent five years from 2011 to 2015. Interestingly, in the last five years (2016-2020), the popularity of assessing reasoning has grown remarkably with 28 studies.

The studies were conducted in 28 different countries on six continents around the world: Europe (n = 31, 49.2%), Asia (n = 20; 31.7%), North America (n = 8; 12.7%), South America (n = 2; 3.2%), Australia (n = 2; 3.2%), and Africa (n = 2; 3.2%). European countries contributed nearly half of the studies (49.2%). The country in which the largest number of studies was conducted was the US (n = 8), followed by China, Germany, Netherlands (n = 7), and Hungary (n = 5), and Finland and Turkey (n = 3).

Across the studies reviewed, there were 41,228 participants ranging from primary school to university. The proportion of test-takers classified relied on grade levels that were tantamount to all age groups. The participants were from the first grade to fifth grades (22.3%), followed by the group of sixth-ninth grade (25.7%) and tenth-twelfth grade (17.9%), while nearly 34.2% of the studies involved a higher education population. However, most of the studies selected (n = 26; 41.3%) conducted in the group of first-fifth grade, followed by the sixth-ninth (n = 22; 34.9%) and the tenth-twelfth grade groups (n = 17; 27.0%). Around a fifth of the reviewed studies (n = 14) targeted university students. In fact, ten studies (n = 10; 15.9%) administered their projects across different grade levels.

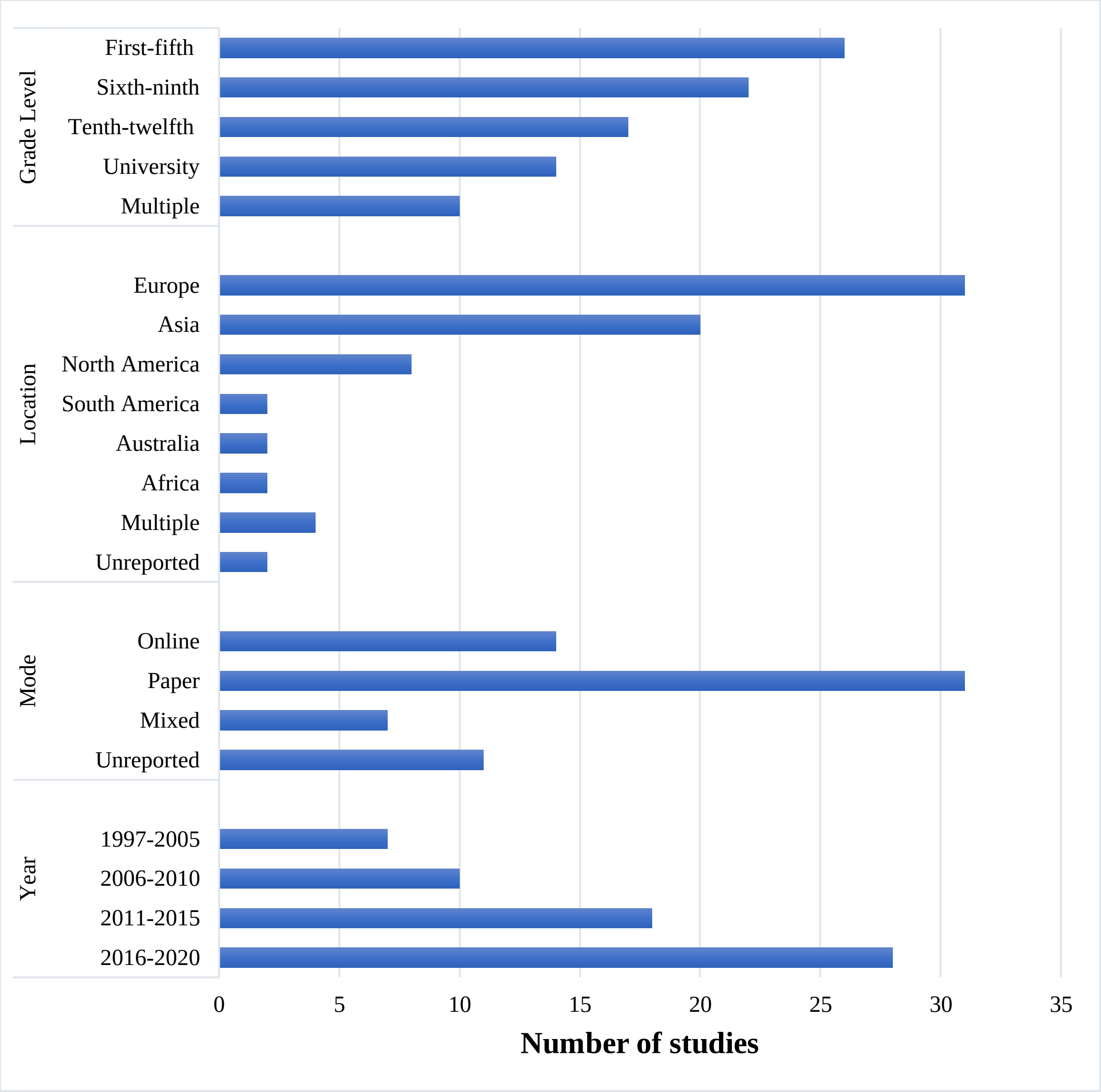


Figure 2.3. Summary of the surface characteristics of the studies reviewed.

For the mode of assessment operation, PP was the most popular administration mode (n = 31; 49.2%) and around 20% (n = 22.2) of the studies made use of an online platform. About 17.5% of the papers did not report the mode of assessment, while approximately 11.1% (n = 7) used both online and PP test instruments.

2.1.6 The construction of instruments for assessing reasoning

As for the combination of form of reasoning and learning domain, 37 of 63 studies reviewed (58.7%) assessed IR across subjects of learning, while 29 of 63 studies (around 46.0%)

measured SR in both specific domains (mathematics, physics, biology, chemistry and social studies) and science as a general domain. Just around 9 (14.3%) studies investigated students' reasoning using a combination of different kinds of reasoning. However, surprisingly, a minority of the studies (under 10%) were conducted with DR tests.

Table 2.2 showed the proportion of IR tasks in accordance with subtest groups mentioned above and the materials constructed elements across the reviewed studies. The results revealed that analogy items have been mostly used in the assessment of IR with 75.7% of studies, followed by series completion and, making up to 64.9% studies applying this kind of task as a main tool to measure IR. Geometric-figure matrix and classification were also frequently found in many instruments in which series completion (56.8%) is more popular than the classification task (24.3%).

Table 2.2. Proportion of tasks and materials in assessing IR (N = 37).

	Problem task				Material			
	AN	SE	CL	GM	Verbal	Number	Geometry	figure
N	28	24	9	21	7	13	29	32
%	75.7	64.9	24.3	56.8	18.9	35.1	78.4	86.5

Note. AN: Analogies, SE: series completion, CL: classification, GE: geometric-figure matrix.

Regarding materials for constructing problems, elements in IR tasks are made of words, numbers, and figures (e.g., geometry, animal, toy, and artifact). The figural object task was the most common one chosen to assess IR abilities of children (Csapó et al., 2014; Csapó et al., 2009; Hamers et al., 1998; Jeotee, 2012; Kambeyo & Wu, 2018; Kyllonen et al., 2019; Muniz et al., 2012; Roberts et al., 2000; Schroeders & Wilhelm, 2010; Stevenson et al., 2013; Van Vo & Csapó, 2020; Wesiak, 2003). Up to 86.5 percent of the selected studies implemented the figure as a main material in assessing IR capacity. Number is also a favourite element selected by developers when designing IR problems. Nearly 35% of the reviewed studies utilized the numeric materials for the main tests in their research (e.g. Bühner et al., 2008; Csapó, 1997; Csapó et al., 2009; Kambeyo & Wu, 2018; Korom et al., 2017; Kyllonen et al., 2019; Nikolov & Csapó, 2018; Van Vo & Csapó, 2020; Wesiak, 2003). The tasks with numeric materials are not only to assess thinking skills, but they are probably also helpful to investigate the participants' mathematical competency in general. The young students tended to perform well or better on simplified abstract items (Figure 2.4) and figural material elements than on contextualized items (e.g. Roberts et al., 2000; Van Vo & Csapó, 2020). Hence, studies used the figural and geometric objects as their main instruments (e.g. Blum et al., 2016; Kyllonen et al., 2019; Muniz et al., 2012; Schroeders & Wilhelm, 2010; Stevenson et al., 2013; Tunteler et al., 2008; Tzuriel & George, 2009; Van Vo & Csapó, 2020; Vogelaar et al., 2019; Wang, 2008), accounting for more than 80%. Although non-verbal tasks were the most common items composed to assess IR (18.9%), verbal elements play a particular role in many learning tasks in school contexts. The reason is that the children's development in scientific concepts does not separate them from the ability to verbally process information (de Koning et al., 2002). Thus, the verbal material was suitable to use in assessment for learning IR in schools. Some studies involved verbal tasks to assess students' IR capacity, such as Kyllonen et al. (2019), Csapó (1997), Csapó et al. (2009), Wesiak (2003), de Koning et al. (2002), Nikolov and Csapó (2018), Vainikainen et al. (2015) and Strobel et al. (2019).

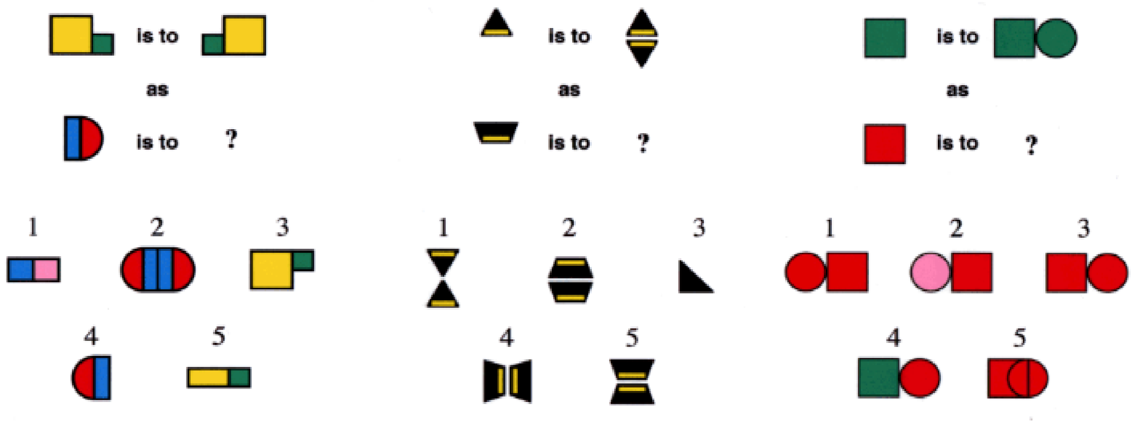


Figure 2.4. Samples of simplified geometric analogy items (from Roberts et al., 2000).

Regarding SR, most the included studies seem to refer to the Lawson Classroom Test of Scientific Reasoning (LCTSR) as a practical framework to adapt and develop their instruments to assess scientific reasoning skills. However, the ratio of tasks or subtests in the instruments may be different depending upon the individual purposes in particular contexts. The control of variables strategy is beyond a SR task since it is considered as a core strategy in learning science, with more than 75% of studies including the tasks in the main tests. Figure 2.5 presents a sample item of SR in control of variables. Physics, chemistry and biology were the most popular school disciplines in which SR studies were conducted.

Regarding item format, multiple-choice structure is still the most popular form of assessing reasoning, but it has become more diverse with visualized presentation in a rich-technology environment, and some testing programs simulated the experiments as real-life phenomena. In spite of the most frequent usage of selected-response format, approximately 90% of studies reviewed, a few studies (under 10%) implemented constructed-response format for the second items of two-tier questions in their SR tests. The constructed-response items were also found in dynamic reasoning tests (e.g. Resing et al., 2017; Vogelaar et al., 2019) which offered more insights into an individual's learning potential than static ones. More detailed technology issues for assessing reasoning will be discussed in the next section.

Ice-cubes and filling level

Timo has an idea. He assumes that ice-cubes melt faster in a large amount of water than in a small amount of water. Which of the following experiments would be a good experiment to test his assumption?

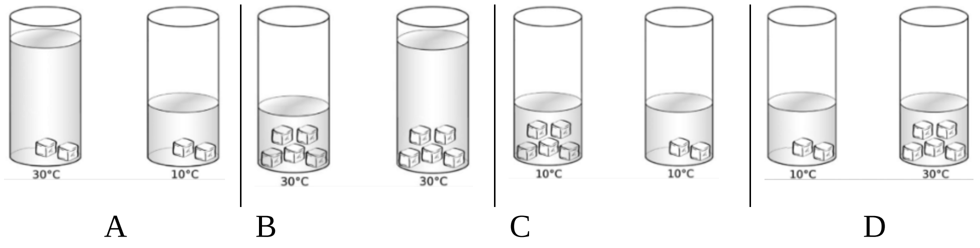


Figure 2.5. A sample of control of variables task (from Schwichow et al., 2016).

To summarize, the most frequent subtests of IR were the verbal and geometric analogies, classifications, number series completions and geometric matrices. The main standardized tests were RPM, Figure Reasoning Test, Vienna Matrices Test, LCTSR and Wason’s four-card task. In fact, RPM and LCTSR are considered as the core rationales for development of IR tests and SR tests, respectively.

2.1.7 Psychometric modeling and technology issues in assessment of reasoning

A psychometric model provides the association between theoretical constructs and empirical assessments. The association between structural quality of an instrument and participants’ response on different items was quantified via psychometric modelling. It allows researchers to estimate and calculate the quality of instruments and empirical data evidence underlying assumptions and predictions. Item response theory (IRT) performs as a helpful technique in the assessment of learning and cognitive potential in educational settings (Stevenson et al., 2013). The quantitative approach was used in most of the studies selected, in which two studies applied a mixed method (qualitative and quantitative). Most studies used a non-experimental approach, and there are some studies applying the quasi-experimental design. Cronbach’s alpha (alpha coefficient) was most often employed to evaluate internal consistency reliability. The primary psychometric property analyses included descriptive statistics, an examination of the internal consistency reliability, validity and then an examination of the correlations between targeted variables under assisting of popular software applications, such as R, SPSS, Mplus, LISREL, Winsteps, ConQuest and Microfact. The popular types of inferential statistics in these studies were descriptive statistics, such as histograms, pirate plots, means and percentages, difference inferential statistics (*t*-test, ANOVA, MANOVA and Mann-Whitney) and associational inferential statistics with Spearman correlation and chi-square. Amongst psychometric approaches, Rasch model measurement seemed to be most frequently used to evaluate the psychometric property of instruments as well as to scale students' performance in the empirical studies selected.

The reasoning tests have been designed and administered in technology-rich environments since 2008. Electronic devices can help to improve cognitive ability testing. It seems that no significant effects of the test media (computers, notebooks, and PP) were determined in terms of psychometric properties of the tests. Some studies investigated the transformation from PP to computer-based testing. Particularly, five studies reported that no significant gender differences were found between the computerized and PP testing (Csapó et al., 2014). The transition from the traditional delivery to an online platform can help to improve the reliability and standardization of the tests (Csapó et al., 2012). For example, studies by Csapó et al. (2009), which assessed IR ability of the 11-year-old children, showed that the paper and online versions were comparable in terms of reliability in both sub-sample and entire sample levels, but the mean scores favoured the paper-based assessment group. Similarly, no media effects were found in the RPM Test (Williams & McCord, 2006), and even the reliability of the online tests was rather higher than that of the paper version (Csapó et al., 2014). Moreover, Schroeders and Wilhelm (2010) conducted a study to assess a verbal, a numerical and a figural reasoning test, which was delivered on computers, notebooks and PP. The authors concluded that only a small and even uncorrelated effect of administration modes were indicated regarding intercorrelation of the structures of the reasoning tests. However, these findings informed that the mean score in the paper format was higher than others. Resing et al, (2017) described that an electronic device is possible to present a series of tasks to track children's responses and the timing of those responses. In computerized simulation assessment, students can operate to control variables and predict the visualized results reasonably. For instance, an empirical study by Al-Balushi and his colleagues (2017) showed that applying three-dimensional animated illustrations at the particulate level of matter allowed students to interact between submicroscopic entities which are not observable. These findings further indicated that rich-visualized environment can enhance learners' reasoning ability.

The most popular online testing platforms were such *Testing Assistée par Ordinateur (TAO)* (Luxembourg) (and it adapted versions), platform Concerto v3.9.14 (Germany) (Blum et al., 2016) and eDia (Hungary) (Csapó & Molnár, 2019).

2.1.8 The effects of gender, age and educational programs on reasoning

Gender differences have presented an important objective when assessing reasoning as indicated in 30% of selected studies. In a comparison of reasoning ability in males and females, inconsistent findings were reported in different studies in both IR and SR. Particularly, some papers reported a significant difference in IR between male and female students, favouring males (Blum et al., 2016; Jeotee, 2012; Kambeyo & Wu, 2018; Kyllonen et al., 2019; Tairab, 2015; Tekkaya & Yenilmez, 2006; Venville & Oliver, 2015). In contrast, another study showed that girls got higher scores than boys on the IR test (Díaz-Morales & Escribano, 2013). Meanwhile, other studies found no significant disparities regarding genders in reasoning ability (Kambeyo, 2018; Molnár, 2011; Salihu et al., 2018; Van Vo & Csapó, 2020), and even no significant gender differences was found in online training program for enhancing basic math content through IR tasks (Mousa & Molnár, 2020). For SR, boys performed better than girls did in some studies (Tairab, 2015; Tekkaya & Yenilmez, 2006; Valanides, 1997), but others found that there was no significant difference between male and female students (e.g. Mayer et al., 2014; Piraksa et al., 2014; Thuneberg et al., 2015). Culture may be an underlying factor impacting these divergent results.

As for the reasoning ability across age or grade levels, around 12.5% of the studies explored the development of reasoning in school-year children. Developments of students' reasoning were observed grade by grade during secondary school years, but the growth rates were different in different pairs of age groups (e.g. Csapó, 1997; Díaz-Morales & Escribano, 2013; Ding, 2018; Edelsbrunner, 2017; Kwon & Lawson, 2000; Molnár et al., 2013; Muniz et al., 2012; Van Vo & Csapó, 2020). In other words, the age-group or grade level was associated to IR ability in children (Stevenson et al., 2013). Particularly, IR has improved across education level system, from the 1st to 5th grade and from the 3rd to 11th grade (Muniz et al., 2012). The most rapid development was identified at the 6th-7th-grade period (Csapó, 1997; Díaz-Morales & Escribano, 2013; Molnár et al., 2013). Likewise, students' scores on SR tests showed a slight increase from first to third grade, followed by a surge in both 4th and 6th grades (Edelsbrunner, 2017) and age range of 13-15 years (Ding, 2018), but the growth slope seemed to slightly reduce after the 9th grade (14 - 15 years) (Kwon & Lawson, 2000). Further, Tairab (2015) found that students from the 8th, 10th and 12th grades performed differently in SR, with the older groups tended to achieve higher scores in control of variables. However, scientific reasoning capacity in university students improved slowly through the entire 4 years in the higher education level (Ding, 2018; Ding et al., 2016). Overall, the patterns of development in reasoning during school showed similar trend across cultures with a noticeable growth during the middle school years (7th to 9th grades).

2.1.9 Reasoning and academic achievement

More than a third of the studies investigated reasoning ability as a significant predictor of academic performance. All the studies concluded that students who achieved higher scores on IR tests tended to receive higher scores in school performance or grade point average (GPA) (Ariës et al., 2016; Csapó, 1997; Díaz-Morales & Escribano, 2013; Mehraj, 2016; Mollohan, 2015; Stamovlasis et al., 2010; Strobel et al., 2019; Tekkaya & Yenilmez, 2006; Venville & Oliver, 2015), with the exception of a study conducted at a Thai university (Jeotee, 2012). Analogical reasoning and SR are the dominant predictors of scientific concept construction and content knowledge (Chuang & She, 2013). The ability and experience to use the prior knowledge was strongly correlated with subtests in the SR tests (proportional reasoning, control of variables and probability reasoning) (Hejnová et al., 2018). Many studies carried out how SR is linked to problem-solving and other competencies. A quarter of the articles reviewed explored the relationships between reasoning and problem-solving. All these findings reported that there was a positive relationship between IR, SR and problem-solving abilities (Boujaoude et al., 2007; Csapó, 1997; Jeotee, 2012; Mayer et al., 2014; Molnár et al., 2013; Rudolph et al., 2018; Vainikainen et al., 2015; Wu & Molnár, 2018).

STEM subjects were the most common disciplines assessed along with reasoning (Boroş & Sas, 2011; Mollohan, 2015; Stender, Venville & Oliver, 2015). Although there was a significant positive association between non-verbal reasoning and performance on mathematics and reading tasks (Salihu et al., 2018), reasoning positively affected learning content knowledge, and vice versa. For example, learning science content from inquiry activities contributed to cultivating children's reasoning proficiencies. A high performance in domain-specific science learning can transfer to domain-general thinking skills, but general cognitive skills such as IR sometimes plays as a mediating role on inquiry skills for the development of children's SR (Stender et al., 2018). Students' GPA (Thuneberg et al., 2015) and mathematics achievement (Stevenson et al., 2013; Valanides, 1997) were significant predictors of students' performance in SR, and SR was the main factor predicting knowledge concept gains (Kwon & Lawson, 2000; Valanides, 1997). Additionally, a strong relationship was found between IR and reading in English as a foreign language in which IR affected stronger than that of socio-economic status (Nikolov & Csapó, 2018). Nonetheless, a study of Hejnová et al. (2018) showed that dimensions of SR did not correlate significantly with school performance either in mathematics and physics. Other findings indicated that there was no or a weak relationship between SR and GPA in higher education (Ding et al., 2016; Jeotee, 2012; Mollohan, 2015).

2.2 Reasoning, motivation, and academic achievement

2.2.1 Science motivation and its role in learning science

Several theoretical perspectives have been developed to explain why students are engaged in an academic activity. The core aspects of educational motivation are students' goals, the intrinsic and extrinsic nature of motivation, students' beliefs about their competencies and students' perceived evaluation of academic tasks. Garcia and Pintrich (1995) recommended that motivation factors include self-perception, effort, intrinsic goal orientation, task value, self-efficacy, test anxiety, self-regulated learning, task orientation and learning strategies. According to Glynn et al., (2009), motivation to learn refers to "the disposition of students to find academic activities relevant and worthwhile and to try to derive the intended benefits from them" (p. 128). Additionally, Anderman and Dawson (2011) summarized four theoretical perspectives on performance of motivation: the goal orientation, socio-cognitive, self-determination and expectancy-value theories. Other scholars have focused on the essential reasons that students choose to engage in tasks, such as mastery goals and performance goals. Meanwhile, socio-cognitive theorists often examine the interactions between the learner, the environment, and other relevant factors. Self-efficacy is one of the main factors in determining a person's beliefs about the ability to complete a task, and students with high self-efficacy are able to execute a learning task regardless of its difficulty (Tuan et al., 2005).

Informed by these different approaches, several instruments have been developed to assess science motivation in an educational setting. Glynn and his colleagues (2009) developed the Science Motivation Questionnaire II, covering five subscales: intrinsic motivation, self-determination, self-efficacy, career motivation, and grade motivation (Glynn et al., 2011). Józsa (2014) proposed the Subject Specific Mastery Motivation Questionnaire, involving six school subjects (reading, mathematics, science, English as a foreign language, art, and music) and school mastery pleasure in 5-point Likert-type items. The students' motivation towards science learning (SMTSL) questionnaire (Tuan et al., 2005) was the first subject-specific assessment tool to assess students' motivation in science learning. The basic theories behind the SMTSL are a combination of constructivist learning and motivation in an educational environment. The questionnaire was composed of 5-point Likert-type questions on self-efficacy, active learning strategies, science learning value, performance goals, achievement goals, and learning environment stimulation.

SM plays an important role in learning science and is a major predictor of science performance in schools (Cavas, 2011; Chan & Norlizah, 2018; Clark et al., 2014; Glynn et al., 2009; Tsai et al., 2015). Several studies (Cavas, 2011; Chan & Norlizah, 2018; Dermitzaki et al., 2013; Tuan et al., 2005) have shown that students with higher motivation are likely to achieve higher performance in learning science, in which self-efficacy is one of the central factors predicting academic achievement at the primary and secondary education levels (Bouffard et al., 2001; Britner, 2008; Peetsma et al., 2005). The 5-year panel analysis study in Korea also found that academic performance and self-efficacy formed a longitudinal causal relationship (Hwang et al., 2016). Additionally, the Programme for International Student Assessment (PISA) reported that 15-year-old students who indicated greater motivation scored higher on PISA school subject tests than their peers within the same country (Mo, 2019).

2.2.2 The effects of age and gender on science motivation

From a socio-cognitive viewpoint, academic motivation is formed from both contextual factors and student cognition (Anderman & Dawson, 2011; Pintrich & Schunk, 2002). Empirical studies (e.g. Dorfman & Fortus, 2019; Józsa et al., 2017) have indicated changing patterns of science motivation across grade levels. The general findings are that students' motivation tends to drop gradually as they move through the school system. A two-year longitudinal study (Bouffard et al., 2001) showed that students' experience declined in terms of self-efficacy belief and learning strategies on reaching secondary school, while Hoffman (2015) demonstrated that the initial passion for formal learning scattered across the middle school years (7th to 9th grades) before plummeting in high school. Furthermore, a 5-year longitudinal study by Gottfried et al. (2001) described academic intrinsic motivation significantly decreasing in a linear trend over the years, but the reduction rates for motivation depended on the particular subject areas, with maths showing the largest decline and social studies seeming unchanged. It appears that the developmental change in students' motivation in schools may be related to biological development.

Concerning gender differences, previous studies on SM showed inconsistent findings in different contexts. No significant gender-related differences have been found in motivation towards science scales (e.g. intrinsic and extrinsic motivation, self-determination and self-efficacy) (Britner, 2008; Glynn et al., 2009; Zeyer, 2010; Zeyer & Wolf, 2010), although boys scored slightly higher points in self-efficacy than girls and girls received higher scores on the self-determination scale (Britner, 2008; Glynn et al., 2009). Some studies that implemented the SMTSL questionnaire found that males did not differ from females in science motivation in terms of self-efficacy, learning environment stimulation or active learning strategies (Andressa et al., 2016; Cavas, 2011; Chan & Norlizah, 2018), but females were significantly more motivated regarding science learning values and achievement goal scales (e.g. Cavas, 2011; Chan & Norlizah, 2018) as well as performance goals (Andressa et al., 2016). Other studies have indicated that females scored higher on achievement goals (King & Ganotice, 2014) and on self-efficacy scales in Earth science (Britner, 2008).

2.2.3 Relationship between reasoning and science motivation

The relationship between intelligence and motivation has been considered in previous studies (e.g. Gagné & St Père, 2001; Preckel et al., 2008; Spinath et al., 2006). For example, the study by Spinath and colleagues (2006) demonstrated that children with higher intelligence are likely to achieve higher academic performance in self-concept, self-efficacy and intrinsic values. A meta-analysis drawing from 74 empirical studies (Kriegbaum et al., 2018) also concluded that intelligence and motivation have a positive relationship (r = 0.17) and estimated 16.6% of the overall explained variance in school attainment. In addition, Chraif and Dumitru (2015) revealed that students who perform better on IR tests tended to obtain higher points on motivation surveys than their peers, and an online assessment by Kambeyo (2018) found a statistically significant correlation between IR and SM though the relationship was not strong.

Moreover, collaborations between parents’ education, cognitive ability and achievement motivation in affecting academic achievement were examined in a study by Ganzach (2000). This study also uncovered that children’s cognitive ability has a positive relationship with their mother’s education but not with their father’s education. A synergistic interaction was also found between cognitive ability and academic motivation in predicting student achievement in schools. In other words, the interplay of reasoning and motivation may be an underlying predictor of school achievement, and both constructs can thus contribute to explaining a higher proportion of explained variance in total.

2.2.4 Parental influences on student motivation and school achievement

Interactive activities in school and family environments are likely to impact children’s academic motivation and achievement (Pintrich & Schunk, 2002). Experiences in classrooms notably impact student motivation and school performance (Bathgate & Schunn, 2017; Hernesniemi et al., 2020). Students’ motivation toward learning is positively tied to parental involvement in schooling (Fan et al., 2012; Gonzalez-DeHass et al., 2005) and parents' education levels (Acharya & Joshi, 2009). There is even a strong link between parental care (warmth and volitional support) and mastery motivation, in which mothers' care was a better predictor of mastery motivation and school achievement (Józsa et al., 2019). Studies on the family-school connection have demonstrated the importance of the relationship between parenting factors and school-related performance (Gonida & Urdan, 2007). As shown in the PISA 2015 results, parental involvement, consisting of parents' participation in school-related activities and parents' interest in their children's school activities, was positively related to not only children's success in learning science, but also to other areas (OECD, 2017a). A longitudinal study by Fan and Williams (2010) also confirmed that there was an aggregate impact of interactive variables, including intrinsic motivation, and parental involvement and engagement on learning maths and English. Moreover, the interactions between parents' education, cognitive ability, and achievement motivation in shaping academic achievement were examined in a study by Ganzach (2000). This supposes that the complex relations between parental involvement in children's schooling and academic motivation are reinforced by the parents' education levels variable which is directly linked to cognitive ability, and all of this has an effect on children's academic performance.

2.3 Studies on comparisons of modes of administration

TBA has been highly beneficial for students, teachers, test administrators and other stakeholders. TBA supports the collection of reliable data, allows personal administration to students during the testing process, and saves time when scoring and analyzing the results. It also permits the direct tracking of students by displaying score reports immediately and storing them in individual logfiles in an integrated data management system (DiCerbo et al., 2017). Despite the many benefits of TBA, its equivalence to traditional assessment is still a subject of debate.

Several studies conducted over the past decade have produced divergent results on the equivalence of TBA and paper-based tests (Gates & Kochan, 2015; Williamson et al., 2017). Research by Kim and Huynh (2010) demonstrated that the English-language test being measured was equivalent in terms of internal consistency across modes of administration and that most items performed similarly between the paper and online groups using differential item functioning (DIF) analysis. As for the reading test, the cross-mode equivalence was confirmed with respect to model construct reliability, and no significant difference was found for multiple-choice format as regards item difficulty (Buerger et al., 2019). However, Schroeders and Wilhelm (2010) found that children achieved higher scores in paper versions than their peers did in computer test versions. They also found that a significant difference occurred in science subjects, but the mean scores in mathematics and social studies were not significantly different between the two modes of administration. Furthermore, a study by Hassler Hallstedt and Ghaderi (2018) concluded that although students achieved lower scores on the tablet-based version of a basic math test than they did on the paper-based one, there were consistent results in terms of validity and reliability across the two test formats. Meanwhile, findings by Neumann and Neumann (2019) also suggested that the construct validity of the tablet-based test version was consistent with that of the paper-based one, but the mean score of the students who took the tablet-based test was higher than that of the group completing the traditional version.

For cognitive tests, the transition from traditional delivery to an online platform may increase reliability and standardization (Csapó et al., 2012). Specifically, the reliability of the inductive reasoning test showed a good level in both the paper and online formats but favored the online mode (Csapó et al., 2014). Moreover, an investigation by Schroeders and Wilhelm (2010) compared the effects of differing media delivery of reasoning tests (verbal, numerical and figural). The authors found no significant differences in test reliability across these different media, but the average score on the paper test was higher than that of the other test. However, a study by Bailey et al. (2018) did not confirm measurement invariance through the structural equation modeling approach across the computer-based and paper-based versions, although the reliability of the spatial test favored the latter. Likewise, Williamson et al. (2017) found that students tend to perform better on the spatial test in online format (effect sizes: d= 0.27 for the Mental Rotation Test), but they performed significantly better on both subtests on the reasoning test in the paper-based version (effect sizes: d= 0.2 and 0.30).

In short, the average score seems to favor the paper-based group, although the psychometric properties of the test, i.e. reliability and validity, are consistent across the two formats. Most previous empirical studies considered evaluating the validity evidence for the internal structure of a test and the proposed interpretation of the mean total scores for particular purposes (Gates & Kochan, 2015; Williamson et al., 2017). It is essential to use multiple sources of evidence to evaluate the performance of a test. This study tends to employ a multifaceted approach focusing on invariance measurement, the performance distribution of the Rasch model scale and DIF analysis to compare equivalence between the two modes of administration.

2.4 Conclusion and discussion

The chapter attempts to pinpoint the possible connections between theoretical concepts and the empirical paradigms to elucidate the construct of the instruments, psychometric approaches in space of assessing reasoning capacity and science motivation in school settings. We also investigated the main relevant factors (demography and cognition) influencing the reasoning ability and science motivation as well as the relationship with academic success of students across the former studies.

Amongst reasoning forms, IR and SR are mostly interested in studying in school settings due to its important roles in learning school subjects and the feasibility of being enhanced in educational contexts (Adey & Csapó, 2012; Klauer & Phye, 2008; Sternberg, 1986; Sternberg & Sternberg, 2012; Wesiak, 2003). IR, the heart of fluid intelligence (Perret, 2015), and SR, one of the two foundational pillars of scientific literacy (Lawson, 2009), are increasingly considered in educational research. The problem tasks for measuring reasoning seemed to remain unchanged over twenty-three years, but most of them have been transformed from PP to TBA versions on computers and other electronic devices. As numerous studies claimed, students should be equipped with well-developed reasoning skills to succeed in 21st-century life. The relevant research has dramatically broadened in recent years in both domain-general and domain-specific thinking, especially applying new technology in assessing thinking proficiency of students.

To assess the ability of IR, a diversity of problem tasks were proposed, but in general they consist of main task, so-called analogies, series completions and matrices varying through verbal, geometric, figural and numerical materials. The analogies tasks, constructed of the materials in geometry and figure, are the most intensively applied in a variety of IR tests. In fact, the young children seem to be more motivated on the IR tasks which are contextualized realistically (e.g., furniture, toys, and animals) than those on artifact objects. However, more studies should continue to investigate the associations among reasoning domains (daily context, domain-specific and domain-free) to provide more empirical evidence for building the appropriate schemas in teaching and assessment purposes in school settings.

In line with a review paper of Opitz et al. (2017), multiple-choice format is still the most popular one to assess IR and SR, but it becomes more diverse in rich-technology media. As transformation into a multimedia environment, the selected-response has been more flexible and interactive in the graphic interface, in which participants can drag and drop the alternative responses to solve the problem tasks. The terms “draggable”, “droppable”, "sortable" and "clickable" have become the most commonly used terminology-related terms in test development within new electronic devices such as smartphones, tablets and wearable devices. Touch screen or drag-and-drop method is a potential approach for both test-takers responding and tracking their processes. Nevertheless, constructed-response format still plays a typical role in particular cases of estimating reasoning ability. For instance, the second item in two-tier questions in constructed-response format is still useful for evaluating the respondents' explanation in SR tasks.

Rasch measurement is a mostly used psychometric modelling chosen in practicing to test models and interpreting the results in reasoning studies. The findings are in line with a review study by Edelsbrunner and Dablander (2019), which exposed that the major psychometric approach for weighing the quality of test items was the infit parameter in Rasch model measurement. Researchers tend to draw interpretations based on not only the infit values for single items to test a good fit model, but also considered classical statistics indices such as discrimination values, percentage of correct answer for inferential purposes. The Wright map is most frequently used to visualize the relationship between participant response and item difficulty. These psychometric approaches could be replicated in investigation of IR and SR. More complex assessments and multivariate analyses are required to describe the relationships between reasoning and other cognitive, non-cognitive, and environmental factors. Systematic modelling structures are called for to establish causal relationships between them.

Gender difference is inconsistent across reviewed studies, but the finding is in line with recent meta-analytic study of Waschl and Burns (2020) for adults, showing that in most situations, effect size was so small, implying no significant gender differences. Other studies revealed that males seem to be favoured on IR tests with figural material. Although this study indicated that the influence of the choice of test battery can cause sex differences, the culture may be one of the silent factors impacting these conflicting results.

Scientific reasoning is often assessed along with IR to examine the relationship between these relevant variables. General thinking in terms of IR and SR are closely tied in predicting scientific concept construction and content knowledge (Chuang & She, 2013). The IR and prior knowledge was main factors in explaining children' SR ability (Hejnová et al., 2018). Thus, assessing the kinds of reasoning contribute to not only understanding the interaction between these kinds of reasoning, but also partly estimating the success of current curricula in enhancing thinking skills for students.

Despite the former studies identified reasoning capacity progressed across the grades with different rates, the results also showed that the reasoning proficiency in children develops most rapidly in the middle school level, recommending that the lower secondary school is the most appropriate time to promote thinking skills. Thus, parents and teachers should be aware of the opportunity to boost children’s reasoning skills in those years. Both physical and social experience influenced scientific reasoning ability (Kwon & Lawson, 2000), so teachers can facilitate more inquiry-based activities through learning science topics to enhance both general cognitive skills and knowledge gains.

CHAPTER 3. METHODOLOGY OF EMPIRICAL STUDIES

3.1 Cross-sectional investigation

An effective developmental study requires longitudinal surveys and takes as long as the developmental process that is being observed. Therefore, developmental programs spanning several years are expensive, require stable long-term funding, and are therefore rare. However, if the environmental conditions for development are constant or change relatively slowly compared to the developmental process under observation, these processes may be estimated through cross-sectional assessment. Though cross-sectional data are more biased if the period covered is longer and the assessed age groups are influenced by different environmental factors (e.g. Baltes, 1968; Maxwell & Cole, 2007), they can provide acceptable estimates of real developmental process when surveying schoolchildren who develop within slowly changing educational systems.

Longitudinal studies of motivation usually take less than 2 years (e.g. 2 months in a study by Datu et al. (2018); 3 months in Schwartz and Waterman (2006); 4 months in Bathgate and Schunn (2017) and 21 months in Chapman (1988)), while studies covering 3 or more years are very rare (e.g. 3 years in Becker et al. (2010); and 8 years in Gottfried et al. (2001)). As for IR, its development has also been studied both in cross-sectional (Csapó, 1997; Molnár et al., 2013; Van Vo & Csapó, 2020) and longitudinal investigations (Ifenthaler & Seel, 2011). Cross-sectional study of the present study conducted in four to five cohorts. Although cross-sectional studies may not provide intact developmental data, the grade-level differences observed may offer a satisfactory estimate of the main trajectories in the Vietnamese context, where the current national education curricula have been consistently implemented across the country since 2002. Table 3.1 provides the brief of three cross-sectional studies in this project.

Table 3.1. A series of cross-sectional studies.

Timeline	Main aims	Instruments	Samples
September & October 2019	- Validating the instruments - Exploring the trajectories of IR and SR - Examining latent factors predicting individuals' IR	- IR test - SMTSL questionnaire - Background questionnaire	5th, 7th, 9th and 11th grades N = 701
October & November 2019	- Exploring developments of IR and SR across cohorts - Investigating the interaction among IR, SR, SM and parental factors in predicting students' STEM achievement	- IR test - SMTSL questionnaire - Background questionnaire	6th, 8th, 10th, and 11th grades N = 733
January & February 2020, 2021	- Developing and validating the CVS test in physics - Investigating development and relationships of CVS and physics motivation in secondary school students - Exploring factors contributing to explaining individual CVS capacity,	- CVS test - SMTSL questionnaire - Background questionnaire	8th, 9th, 10th, 11th, and 12th grades N = 807

3.2 Instruments

3.2.1 Reasoning tests

3.2.1.1 Inductive reasoning test

Klauer and Phye (2008) pointed out that IR refers to discovering regularities by finding similarities, dissimilarities, or a combination of both with regard to the attributes of or relations to or between objects. Inspired by the point, the IR tasks were composed of domain-general and non-verbal material with 4 subtests: figure analogies (FA), figure series completion (FS), number analogies (AN) and number series completion (NS). The item pool was developed by the Research Group on the Development of Competencies at the University of Szeged. The original items included four subtests in Hungarian translated into English and other languages. Several empirical studies were conducted with school-age populations to establish its reliability and predictive validity across cultures, such as in in Hungary (Csapó, 1997; Pásztor, 2016; Pásztor et al, 2017) and Finland (Csapó et al., 2019; Molnár & Csapó, 2011), Namibia (Kambeyo, 2018), China (Wu & Molnár, 2018) and Indonesia (Saleh & Molnár, 2018).

The foundational criteria for selecting items were based on the structure of each item and evidence from previous studies. Firstly, we investigated the composition and the rules in each item because one item typically contains certain kinds of figure or numbers. We were concerned about the diversity of the items and avoided tasks which contained similar rules or constructions of the test. Then, we referenced the item difficulty from empirical studies. The test was expected to measure the appropriate abilities covering all students in the sample. Finally, the 32-item IR test was adapted and translated from English into Vietnamese, with a subtest consisting of eight items. A correct answer was assigned 1 point, and an incorrect answer was assigned 0 points for all the items.

The online test was developed within the Electronic Diagnostic Assessment System (eDia), a platform created by the Center for Research on Learning and Instruction at the University of Szeged (Csapó & Molnár, 2019; Molnár & Csapó, 2019). The eDia platform supports the entire item writing and test editing process as well as delivering tests and providing feedback. The eDia system is an easy-to-use diagnostic instrument containing item banks to support personalized teaching and learning in reading, mathematics and science. Students may access eDia using a standard Internet browser, such as Mozilla Firefox and Google Chrome, with the IT infrastructure available at schools (desktop computers and mobile devices). Beyond its main function, the online platform may be used for completing sophisticated research tasks, such as logfile analysis (Greiff et al., 2018; Csapó & Molnár, 2019).

Figure 3.1. presents examples of IR tasks available on the eDia platform. The exact same layout was used in the PP version (See Appendix B).

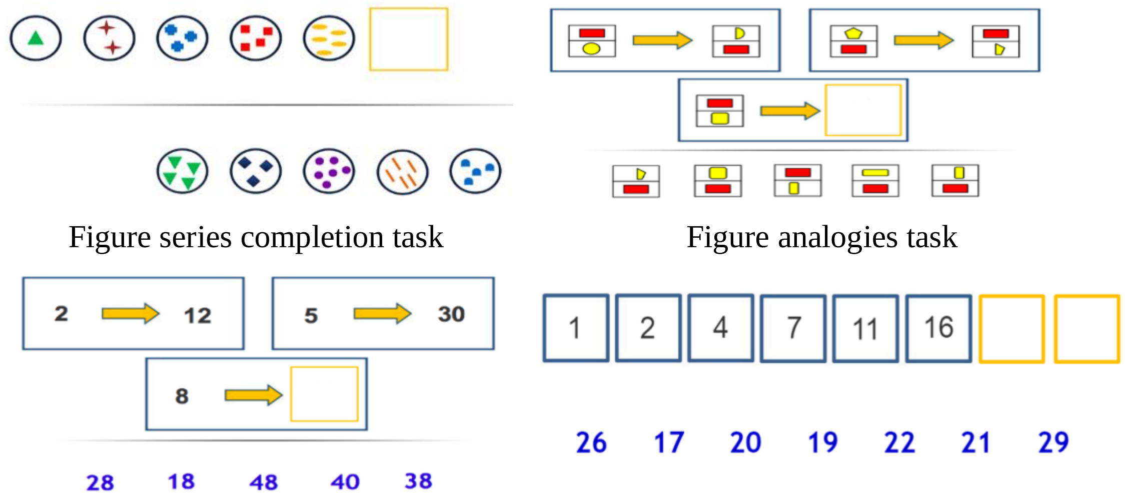


Figure series completion task

Figure analogies task

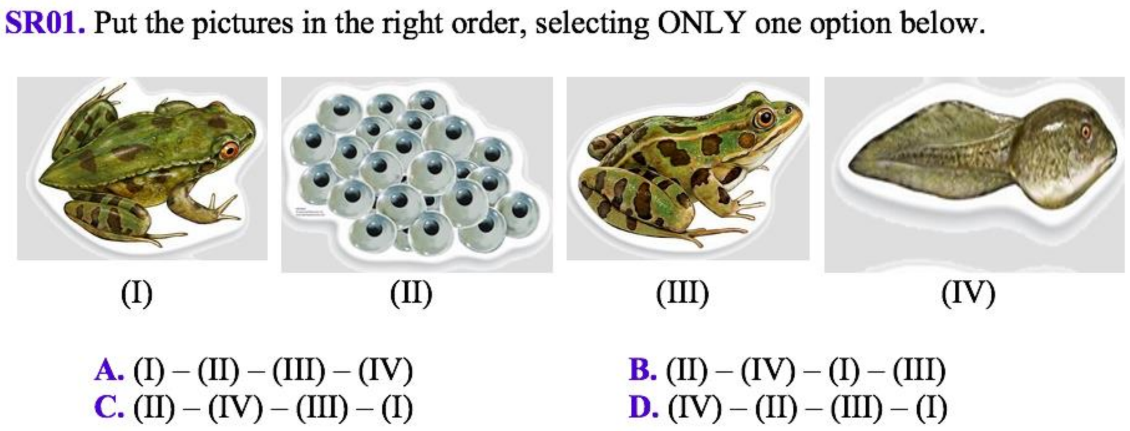
Number analogies task

Number series completion task

Figure 3.1. Examples of items on the inductive reasoning test.

3.2.1.2 Scientific reasoning test

We composed 18 items to measure SR, covering main tasks such as conservation, classification, proportional reasoning, correlational reasoning and IR with elements constructed of science content. The content knowledge of the test items is related to basic concepts of science subjects in secondary educational curricula in Vietnam. Most of the items were adapted and translated into Vietnamese from the original test of Korom et al. (2017). Two items were adapted from LCTSR (Lawson, 2000) and the scientific reasoning test by Hanson (2016), and only one new item was composed by the authors. The items were modified into multiple-choice format and each item contained a correct answer and three distractors. We used a multiple-choice format because it is the most popular one in measuring SR (Opitz et al., 2017), and it can provide the opportunity to increase the precision when measuring a larger number of respondents with smaller effect sizes than other question formats (Schwichow, Croker, et al., 2016). We also tried to minimize the impact of students' reading ability levels by reducing the texts and presenting more visualized representations with figures and graphs. The Figure 3.2 illustrates two items in the SR test (see Appendix C).

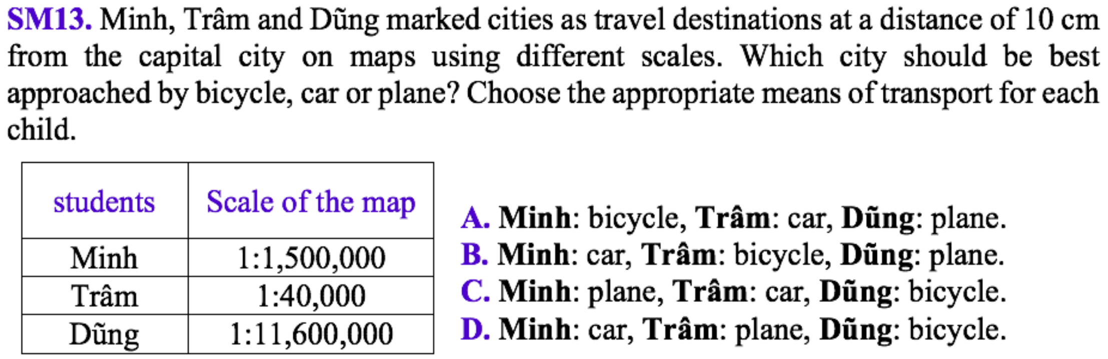


SR01. Put the pictures in the right order, selecting ONLY one option below.

(I) (II) (III) (IV)

A. (I) – (II) – (III) – (IV) **B.** (II) – (IV) – (I) – (III)
C. (II) – (IV) – (III) – (I) **D.** (IV) – (II) – (III) – (I)

a) A series completion with discipline content task.



SM13. Minh, Trâm and Dũng marked cities as travel destinations at a distance of 10 cm from the capital city on maps using different scales. Which city should be best approached by bicycle, car or plane? Choose the appropriate means of transport for each child.

students	Scale of the map
Minh	1:1,500,000
Trâm	1:40,000
Dũng	1:11,600,000

A. Minh: bicycle, **Trâm**: car, **Dũng**: plane.
B. Minh: car, **Trâm**: bicycle, **Dũng**: plane.
C. Minh: plane, **Trâm**: car, **Dũng**: bicycle.
D. Minh: car, **Trâm**: plane, **Dũng**: bicycle.

b) A proportional reasoning task.

Figure 3.2. Examples of (a) a series completion with discipline content task and (b) a proportional reasoning task in the SR test.

3.2.1.3 Control of variables strategy in physics test

The control of variables in physics (CVSP) test involves 24 items, including three subskills of CVS: Identifying controlled experiments, interpreting the outcome of a controlled experiment and understanding the determinacy of confounded experiments. The content of test items is related to physics including basic concepts matching to secondary educational curricular in Vietnam such as mechanics, heat and thermodynamics, and electricity and electromagnetism. The 24-item test covers both cognitive processes in terms of CVS and domain-content in physics.

We developed the test items in the multiple-choice format, comprising a stem with three distractors and one correct answer. A correct answer and incorrect answer were scored 1 and 0 points, respectively. The test specification and development are discussed further in Section 6.3.1.

3.2.2 Student questionnaire

3.2.2.1 Students' motivation toward science learning questionnaire

Our study adapted the SMTSL questionnaire developed by Tuan, Chin and Shieh in 2005. Tuan et al. (2005) supposed that self-efficacy, science learning value, learning strategies, individual learning goals and learning environment stimulation are foundational elements of motivational factors in assessing students' science learning motivation. The questionnaire was therefore developed based on these components. The *self-efficacy* subscale measures students' beliefs about their capacity to perform well on science learning tasks (e.g., I am sure that I can do well on science tests). The *active learning strategies* subscale focuses on how students apply the various strategies to acquire new knowledge based on their personal experience (e.g., When learning new science concepts, I connect them to my previous experiences). *Science learning value* refers to what students can achieve in their daily lives when they attend science courses, such as scientific reasoning, problem-solving skills, and science knowledge (e.g., I think that learning science is important because it stimulates my thinking). *Performance goals* denote students' goals in science learning such that they compete with other students and draw their teacher's attention. The *achievement goals* subscale assesses students' satisfaction with their performance in science learning (e.g., During a science course, I feel most fulfilled when I am able to solve a difficult problem). In the current study, our adapted instrument focused on achievement goals. The performance goals subscale was not included in this study because we reduced the number of items to match the limited time of the 45-minute period suggested by principals at the participating schools. The *learning environment stimulation* subscale entails the curriculum, instruction, and the learning environment, which influences students' science learning motivation (e.g., I am willing to participate in this science course because the teacher does not put a lot of pressure on me).

Various empirical studies (e.g. Cavas, 2011; Chan & Norlizah, 2018; Dermitzaki et al., 2013; Shaakumeni & Csapó, 2018; Tuan et al., 2005) have provided evidence that the instrument we have selected is reliable and valid in cross-cultural contexts. For example, Cronbach's alpha for the entire questionnaire in Taiwan was 0.89 and ranged from 0.70 to 0.89 for each subscale (Tuan et al., 2005), it was 0.87 in Turkey, ranging from 0.54 to 0.85 for the scales (Cavas, 2011), and it was 0.68 to 0.82 in Greece (Dermitzaki et al., 2013), 0.84 in Malaysia (Chan & Norlizah, 2018), and 0.79, ranging from 0.66 to 0.77, in Namibia (Shaakumeni & Csapó, 2018). In this study, we employed the adapted questionnaire and analysed 18 items on five subscales: self-efficacy, active learning strategies, science learning value, achievement goals and learning environment stimulation.

3.2.2.2 Students' motivation and attitude in learning physics questionnaire

The questionnaire was adapted and translated into Vietnamese languge from the student questionnaire in TIMSS 2015 (separate science subjects: Physics in school) (Hooper et al., 2013; TIMSS, 2015). The questionnaire involves three main motivational scales as follows: students like learning physics, students' views on engaging in physic lessons, and students' confidence in learning physics.

The students like learning physics (hereafter referred as "Like learning" or "LL") scale measures how much students like learning physics in the school based on their level of agreement with the nine statements. In this scale, students responded to the following question "How much do you agree with these statements about learning physics?" with nine statements (e.g., "I enjoy learning physics", "I wish I did not have to study physics", "I learn many interesting things in physics").

The students' views on engaging teaching in physic lessons (hereafter referred as "Engaging teaching" or "ET") scale was used to ask students' perspectives on the physic lessons in schools. The scale starts with the question "How much do you agree with these statements about your physics lessons?" with the following ten items (e.g., "I know what my teacher expects me to do", "My teacher is easy to understand", "I am interested in what my teacher says").

The students' confidence in learning physics scale (hereafter referred as "Confidence in learning" or "CL") was used to estimate how confident students feel about their ability in physics. It was composed of the question "How much do you agree with these statements about physics?", followed by the possible levels of agreement along with the nine statements (e.g., "I usually do well in physics", "I learn things quickly in physics", "I am good at working out difficult physics problems"). There are four possible options ("Agree a lot", "Agree a little", "Disagree a little", and "Disagree a lot") to respond to each statement.

3.2.2.3 Background questionnaire

We adapted the student questionnaire from PISA 2015 (OECD, 2015) and translated into Vietnamese language. The questionnaire was designed to collect data related to students' background, parents' education level and parental involvement in schooling. Students were asked to respond to items by weighing their perceptions about parent's involvement regarding parental support, engagement, and interests in school activities. In this part, students also reported their scores of the discipline tests in the previous semester. The background questionnaire is placed in the first section of each test instrument. The questionnaire was embedded into both PP and online versions.

3.3 Data analysis

3.3.1 Factor analysis

Principal component analysis (PCA) is a popular statistical method for reducing data with several dimensions or variables. Such method uses linear combinations of the variables to project the data with fewer dimensions. PCA is a classical multivariate statistical method that is used to interpret the variation in high-dimensional interrelated dataset. PCA reduces the high-dimensional interrelated data to low-dimension by linearly transforming the old variable into a new set of uncorrelated variables called principal components when retaining the most possible variation. The eigenvalues can help to retain the number of principal components. In general, principal components with eigenvalues > 1 contribute greater variance and should be retained for further analysis. Using the scree plot as the structure to analyse the criterion found, and factor loadings of the items is at least above .30.

Confirmatory factor analysis (CFA) is one of the appropriate methods to evaluate the construct validity of the adapted instruments with the Structural Equation Modelling (SEM) approach. We utilized CFA to evaluate the construct validity of the adapted instruments as a criterion for further analysis with the Rasch model. CFA is often conducted to assess the fit of the model using the standardized root mean squared residual (SRMR), the root mean square error of approximation (RMSEA), the comparative fit index (CFI), and the Tucker–Lewis index (TLI). Fit indices refer to a measure of how well the rotated matrix matches the original matrix. It involves goodness-of-fit statistics (TLI, CFI), which require large values and compares a reproduced correlation matrix to a real correlation matrix. Residual statistics (RMSEA, SRMR), which are expected to have small values, look at the residual matrix. The RMSEA is also used to calculate the "probability of a close fit" or p-close statistic. The CFI assesses the relative improvement in the fit of a model compared with the baseline model, and ranges between 0 and 1. Other index is weighted root mean square residual (WRMR), which was proposed as being suitable with non-normal outcomes. A model is accepted when the WRMR value is equal to or less than 1.0 for the model (Yu, 2002). The popular cut-off criterion is used mostly to assess the model fit in educational research: RMSEA < 0.06, CFI > 0.90, SRMR < 0.08 (Hu & Bentler, 1999). Confirmatory factor analysis was also performed to investigate the equivalence of the factor structure of the test across the administration modes.

3.3.2 Rasch model measurement

IRT is an encouraging technique for data analysis in assessing learning and cognitive potential in educational settings (Stevenson et al., 2013). Rasch model is a psychometric modelling measurement chosen for investigating test models and interpreting the results. The Rasch model, named after George Rasch, considers the relationship between the type of error made and the ability of the student during the assessment. Rasch's approach was used to examine the relationship between the test-takers' abilities and their responses to the test items. This psychometric model presents the probability of a person solving an item predicted from the relationship between the person's ability level in describing a latent trait and the item's difficulty level on the same continuum (Andrich & Marais, 2019). A "logit" scale was applied to express item difficulty on a linear scale that can essentially broaden from negative infinity to positive infinity. This model plays an important role in test development progression. The model Rasch model can be used to scale dichotomous and polytomous items (Partial Credit analysis).

As reviewed from a study by Edelsbrunner and Dablander (2019), the major psychometric approach in assessment of scientific reasoning is Rasch modelling with a focus on drawing interpretations based on both item infit statistics and classical statistics (e.g. discrimination values and percentage of correct answer). Therefore, in the present study, we employed the Rasch model in ACER ConQuest software with dichotomous items (Adams & August, 2010) to convert the raw score data to a linear scale. Such outputs of the program provide the main parameters of model fit statistics and reliability, item-model fit, item ability distribution fit, and some main classical statistics indices.

Regarding dimensional model fit, the efficacy of the possible models in Rasch model analysis is often based on three indices: the deviance, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). The deviance is –2 log(L), with L being the maximum of the likelihood function given the model. The AIC and BIC are calculated from the deviance: $AIC = \ deviance \ + \ 2 \, n_p$ and $BIC = \ deviance \ + \ log(N) \, n_p$ with n_p being the number of parameters and N being the number of persons (Wilson et al., 2008). By comparing the final deviances, AIC and BIC of the models, a model that gets the lower coefficients implies a better fit. In each model, the cut-off standard for the acceptable value of the infit index is in the range of 0.77 - 1.30 (Griffin, 2010). An item is considered as an infit one in the Rasch measurement model if it meets this criterion. A good item is denoted that the probability of students who have a correct answer for that item will increase with their proficiencies. A Wright map is the most visualized means to explore the relation between participants' response and item difficulty.

3.3.3 Differential item functioning and differential bundle functioning

The differential item functioning (DIF) analysis was utilized to examine statistical characteristics of an item. DIF examination provides evidence about functions by comparing different abilities for members of individual groups to detect whether an item is fair or not among the groups. This approach can be used to compare students' performances between subgroups such as gender and administration modes at the item level.

In the dissertation, we employed DIF analysis in the dichotomously scored items via Angoff's delta plot method (see Angoff, 1982) with the R deltaPlotR package (Magis & Facon, 2014). The main assets of the approach are that it is a simple and straightforward method, which relies on the particular items themselves, does not require intensive addition of other power indices, so it can easily handle with small samples of respondents without relying on asymptotic assumptions. The perpendicular distances of all delta points to the major axis are calculated. If an item is located above the major axis (large positive distances), it is estimated to be easier for respondents in the reference group, and vice versa. A small distance implies that the item difficulty is similar across groups, indicating as a no DIF item. To test measurement invariance for polytomous items, we used the R lordif package (Choi et al., 2011) after testing the presence of DIF under the logistic regression framework. Pseudo R^2 statistics was suggested as magnitude measures and classified DIF as negligible (< 0.13), moderate (between 0.13 and 0.26), and large (> 0.26).

Furthermore, we also conducted differential bundle functioning (DBF) or item bundle DIF analysis which is the natural extension of the DIF approach. The framework for a studied DBF often includes a bundle of items connected with the corresponding content area that matches the test specifications and the associated subtest consisting of the remaining items. In the current study, we applied the framework of Douglas et al. (1996), which used the Simultaneous Item Bias Test (SIBTEST) method (see Shealy & Stout, 1993) to examine DIF in bundles. Douglas, Roussos, and Stout (1996) proposed a multidimensionality-based DIF analysis paradigm including DBF. The basic principle for a recognition of a differential functioning bundles is that those items are identified through a common secondary ability in addition to the target items. The R mirt package (Chalmers, 2012) was applied to assess DBF within the function of the SIBTEST. The SIBTEST provides an estimate of the unidirectional DIF index (β_s). A negative β_s indicates a DBF favouring the focal group and a positive β_s implies a DBF favoring the reference group (Douglas et al., 1996; Shealy & Stout, 1993).

3.3.4 The symmetric log-logistic model

There are several forms of the sigmoid curves to simulate the developmental process in nonlinear regression approach. One of the most common curves is the symmetric log-logistic model as follows:

$$y = c + \frac{d - c}{1+exp(b(log(x) - log(ED50)))} \tag{3.1}$$

Where:

y: the response,

c: the lower limit of the response when the dose x approaches infinity,

d: the upper limit when the dose x approaches 0,

b: the slope around the point of inflection,

ED50: the dose required to reduce the response half-way between the upper and lower limit.

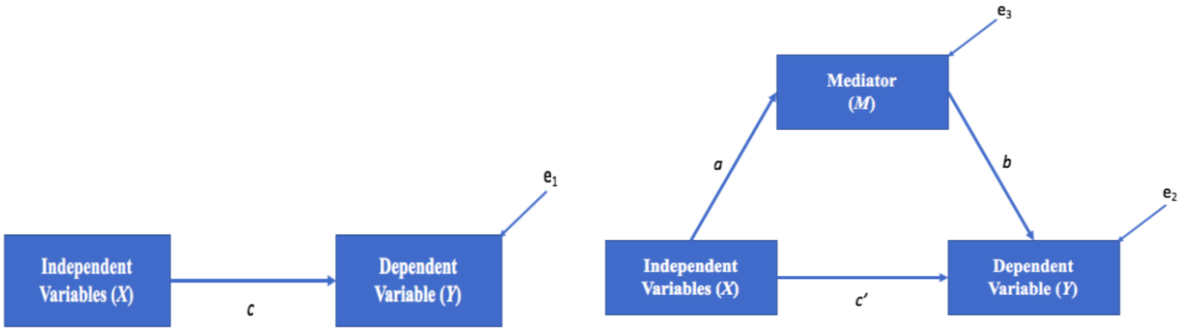
The parameters of a sigmoid curve can be tailored to R drc package (Ritz et al., 2015). We manipulated the feature of the development by fitting the scores with a logistic function in Equation 3.1, in which x and y represent grade levels and student scores, respectively. In the logistic model, to identify the grade level where the most rapid change happens in students' scores, we computed the derivative of the logistic equation at x = ED50.

3.3.5 Bayesian model averaging

To examine the better predictors, performing Bayesian estimation analysis with Bayesian model averaging (BMA) is often used to identify these potential models. The approach can potentially indicates significant improvements in comparison with the existing methods in terms of both predictive and explanatory ability (Genell et al., 2010; Hair et al., 2010). First, in the BMA technique all the best possible models were evaluated according to model fit measured by the Bayesian information criterion (BIC) based on posterior model probabilities. Then, for each explanatory variable the posterior effect probability was computed by averaging over the posterior model probabilities of all models fit. Finally, the average mean and standard deviation of each regression coefficient was estimated by weighted averaging of coefficients under each separate model (see Genell et al., 2010; Raftery et al., 2020). In the current project, we employed the BMA analysis with linear regression models in R package BMA (Raftery et al., 2020). The best models were recommended based on the BIC for each manipulation.

3.3.6 Path analysis

Beyond the estimation of causal effects of the certain variables, understanding how these variables exert their influence on the outcomes is important in analysing causal mechanisms in applied research. In general, this method involves a set of linear regression models fitted and the estimates of "mediation effects" computed from the fitted models in the SEM approach. The single mediator model ($X \to M \to Y$) is illustrated in Equation 3.2 to 3.4 and Figure 3.4 (MacKinnon et al., 2012):



a) a direct effect.

b) a mediating effect.

Figure 3.3. Path analysis with direct and indirect effects.

$$Y = i_1 + c\,X + e_1, \tag{3.2}$$

$$Y = i_2 + c'\,X + b\,M + e_2, \tag{3.3}$$

$$Y = i_3 + a\,X + e_3, \tag{3.4}$$

Where:

X is the independent variable,

Y is the dependent variable,

M is the mediator,

$i_1, i_2,\ and\ i_3$ are intercepts,

$e_1, e_2,\ and\ e_3$ are residuals.

Equation 1 presents the coefficient c representing the total effect (Figure 3.4a). In Equation 2, the coefficient c' denotes the relation between *X* and *Y* controlling for *M*, representing the direct effect. The coefficient b denotes the relation between *M* and *Y* controlling for *X*. In Equation 3, the coefficient *a* indicates the relation between *X* and *M*. Equations 3.3 and 3.4 are represented in Figure 3.4, demonstrating a single mediator model, which depicts how the total effect of *X* on *Y* is apportioned into a direct effect relating *X* to *Y* (path *c'*) and an indirect effect on *Y* through a mediated effect of *M*. Path *a* represents the effect of *X* on the proposed mediator, while path *b* is the effect of mediator *M* on *Y* that is partially out the effect of *X*. The total effect of *X* on *Y* can be expressed as the sum of the direct and indirect effects: $c\ =\ c'\ + ab$ (Preacher & Hayes, 2008). The Sobel test is frequently used to test whether a mediator carries the influence of an independent variable to a dependent variable. R, SPSS, Mplus and SAS programs are available packages, functions or macros plus for mediation effects. In the present studies, we implemented either lavaan package (Rosseel, 2012) and Mplus 7 (Muthén & Muthén, 2012) to employ the path analyses in the mediator models.

Other popular statistical tests such as *t*-test, analysis of variance (ANOVA) and Tukey's honestly significant difference (Tukey's HSD) test were used to examine differences between subsamples. In the case, we utilized the R psych package (Revelle, 2019) and ggplot2 package (Wickham, 2016) to calculate and visualize the findings.

CHAPTER 4. EXPLORING INDUCTIVE REASONING AND STUDENTS' MOTIVATION TOWARD SCIENCE LEARNING ACROSS GRADE LEVELS

4.1 Introduction

IR is one of the primary mental abilities of the fluid intelligence (Kinshuk et al., 2006), and the most broadly researched amongst cognitive skills. Numerous empirical studies investigated the relationship between IR ability and academic performance (e.g. Ariës et al., 2016; Csapó, 1997; Díaz-Morales & Escribano, 2013; Mehraj, 2016; Mollohan, 2015; Stamovlasis et al., 2010; Strobel et al., 2019; Venville & Oliver, 2015). The previous studies showed that IR was one of the main predictors of school performance.

The importance of motivation in learning has also been examined in several studies. Most studies have determined that positive motivation in learning not only promotes students' academic performance during their school years, but is also one of the main prerequisites for their success in future (e.g. Duckworth et al., 2011; Eccles & Wigfield, 2002; OECD, 2017b). The former studies (e.g. Glynn et al., 2011; Chan & Norlizah, 2018; Dermitzaki et al., 2013; Glynn et al., 2009, 2011; Kambeyo, 2018) found that SM and science achievement are closely tied regardless of education levels. Therefore, enhancing SM has been an important goal of formal education and contributed to promoting science literacy for tomorrow's citizens.

Assessing IR and SM is likely to provide useful information for educational reform processes, policymakers and teachers. Particularly, a broadly accessible measurement data from a longitudinal (e.g. Hwang et al., 2016; Robinson et al., 2019) and cross-sectional assessment (e.g. Dorfman & Fortus, 2019; Józsa et al., 2017) can offer a deeper understanding insight into effective factors of IR and SM. It may be meaningful to estimate the efficacy of current curricula in terms of enhancing IR ability and supporting students' motivation as well as proposing improved programme practice in schools in the future.

Moreover, researchers (e.g. Anderman & Dawson, 2011; Pintrich & Schunk, 2002) have agreed that children's academic motivation is as a result of relative dynamic factors of both dispositional and contextual variables. Some components of motivation arise from individual characteristics, and others emerge from direct and indirect interactions in families, schools and society. The existing studies discussed associations between students' motivation with common understudied factors, such as age (e.g. Heckhausen et al., 2010), gender (e.g. Britner, 2008; Cavas, 2011; Chan & Norlizah, 2018), intelligence (e.g. Kriegbaum et al, 2018) and parental involvement (e.g. Fan et al., 2012) and academic achievement (Ganzach, 2000; Gonida & Urdan, 2007). IR is also closely tied to science achievement (e.g. Csapó, 1997; Díaz-Morales & Escribano, 2013; Mehraj, 2016; Mollohan, 2015; Venville & Oliver, 2015). Whether IR and SM establish an interactive relationship in predicting science performance is still an unexplored question.

This cross-sectional study therefore attempts to fill the gaps in the literature by exploring patterns of IR, SM across grade cohorts. We also investigate the relationships between IR, SM, grade level, parental factors and other relevant variables. Our efforts are expected to draw a partial developmental pattern of IR and SM among children and relevant factors explaining individuals' IR and SM in Vietnamese context.

4.2 Research questions and hypotheses

The study aims to evaluate the adapted instrument, investigate the students' performances in IR and SM in different grade cohorts. We also identify dominant factors affected by IR and SM among individuals in school contexts. Therefore, the adapted test instruments were employed to answer the following research questions and hypotheses:

RQ1.1: What is the evidence for the validity and reliability of the adapted instruments (IR test and SMTSL questionnaire) in Vietnamese context?

H1.1: It is expected that the psychometric properties are acceptable for the IR test (e.g. Csapó et al., 2019; Kambeyo, 2018; Molnár & Csapó, 2011; Wu & Molnár, 2018) and SMTL questionnaire (e.g. Cavas, 2011; Chan & Norlizah, 2018; Dermitzaki et al., 2013; Shaakumeni & Csapó, 2018; Tuan et al., 2005).

RQ1.2: How are the IR capacities and SM different among grade cohorts?

H1.2: We hypothesized that students in upper grade perform better than did their counterparts in the lower grade on the IR test (Csapó, 1997; Díaz-Morales & Escribano, 2013; Molnár & Csapó, 2011; Molnár et al., 2013). Students' motivation is assumed just have a slight decrease through grade levels (e.g. Dorfman & Fortus, 2019; Józsa et al., 2017).

RQ1.3: Is there a significant difference between boys and girls on the IR test and the SMTSL questionnaire?

H1.3: No gender difference between boys and girls is expected to observe on the IR test (Jeotee, 2012; Kambeyo, 2018; Molnár, 2011; Tekkaya & Yenilmez, 2006; Venville & Oliver, 2015), but females are supposed to be significantly more motivated towards science learning than males (e.g. Cavas, 2011; Chan & Norlizah, 2018; King & Ganotice, 2014).

RQ1.4: Which factors contribute to individual IR abilities among students?

H1.4: We assume that grade level (school-age group), school performance, father's education level, mother's education level are latent predictors of individual IR ability.

RQ1.5: To what extent is IR ability related to SM?

H1.5: It is expected that a close relationship is established between IR and SM (Cavas, 2011; Hu et al., 2016; Tee et al., 2018).

RQ1.6: Which factors contribute to explaining individual SM among students?

H1.6: We assume that IR, student age (or grade level), gender, science achievement, mother's education level, father's education level and parental involvement are the main underlying factors predicting students' SM.

4.3 Method

4.3.1 Participants

The study assessed 701 students at public secondary schools in An Giang Province (Vietnam). We attempted to select schools that can represent the quality of the students in the area. We selected 20 classes randomly from six schools in total, and students in the intact class participated in this study. As presented in Table 4.1, the properties of four cohorts in the research involved over 130 students for each grade group with the distribution of students by grade levels and demographic characteristics. The study was conducted in August and September 2019. Students spent 45 min to complete the test instrument under examination conditions. We administered the test instrument as part of the regular school timetable. Students took the online test in their school computer labs, and others did the test in PP mode in their regular classrooms (see more testing equivalence at Section 7.6).

Table 4.1. The study samples.

Grade	N	Male/female ratio (%)	Mean age (years)	Age range (Years)	No. of classes
5	157	49.7/50.3	10.3	9.8 - 10.8	4
7	222	48.2/51.8	12.2	11.8 - 13.3	6
9	135	42.2/57.8	14.1	13.8 - 14.6	5
11	187	48.1/51.9	16.2	14.8 - 15.5	5
Total	701	47.4/52.6	13.2	9.8 - 15.5	20

4.3.2 Instruments

Instrument of this study involves three parts (background, SMTSL questionnaire and IR test). The background part was adapted from the student survey for PISA 2015 (OECD, 2016) (see Section 3.2.2.3). The science motivation questionnaire consisted of five subscales: self-efficacy, active learning strategies, science learning value, achievement goals and learning environment stimulation (see Section 3.2.2.1). The 32-item IR test was adapted, which includes four tasks: figure series completion, figure analogies, number series, and number analogies (see Section 3.2.1.1).

Two experts and three secondary school teachers supported to review the adapted test instruments. The final version of the test instruments was implemented to field testing as a pilot study with ten students in two different schools. These students had similar traits with our proposed study sample. Then, each of the items in the pilot study was discussed and modified again before the data collection for the main study took place.

4.3.3 Procedure and data analysis

For the paper-based assessment, the students were given a test booklet containing a questionnaire section and an answer sheet. The teachers guided the students through the appropriate practice items step by step following our procedures in their daily classrooms under the supervision of their teachers. For online administration, the students accessed the eDia platform and registered for the online instrument with a personal code (see Csapó & Molnár, 2019). The students completed the test and questionnaire in the schools' labs with computers or other electronic devices (iPad or smartphone). Two teachers observed and provided technology support in the computer rooms.

CFA was employed to test the model to empirical data with Mplus 7 (Muthén & Muthén, 2012) before doing further analyses with Rasch model. The data were analysed by the Rasch model in ACER ConQuest software for polytomous items with partial credit model (PCM) analysis for the SMTSL questionnaire (Adams & Wu, 2010) and dichotomous items for the inductive reasoning test (Adams & August, 2010). The results were scaled in maximum likelihood estimation (MLE) in Rasch model measurement. All the data sets further were manipulated in R software application version 3.5.3 (R Core Team, 2019), including the common package libraries such as psych (Revelle, 2019), yarrr (Phillips, 2017) and sciplot (Morales, 2020).

4.4 Results

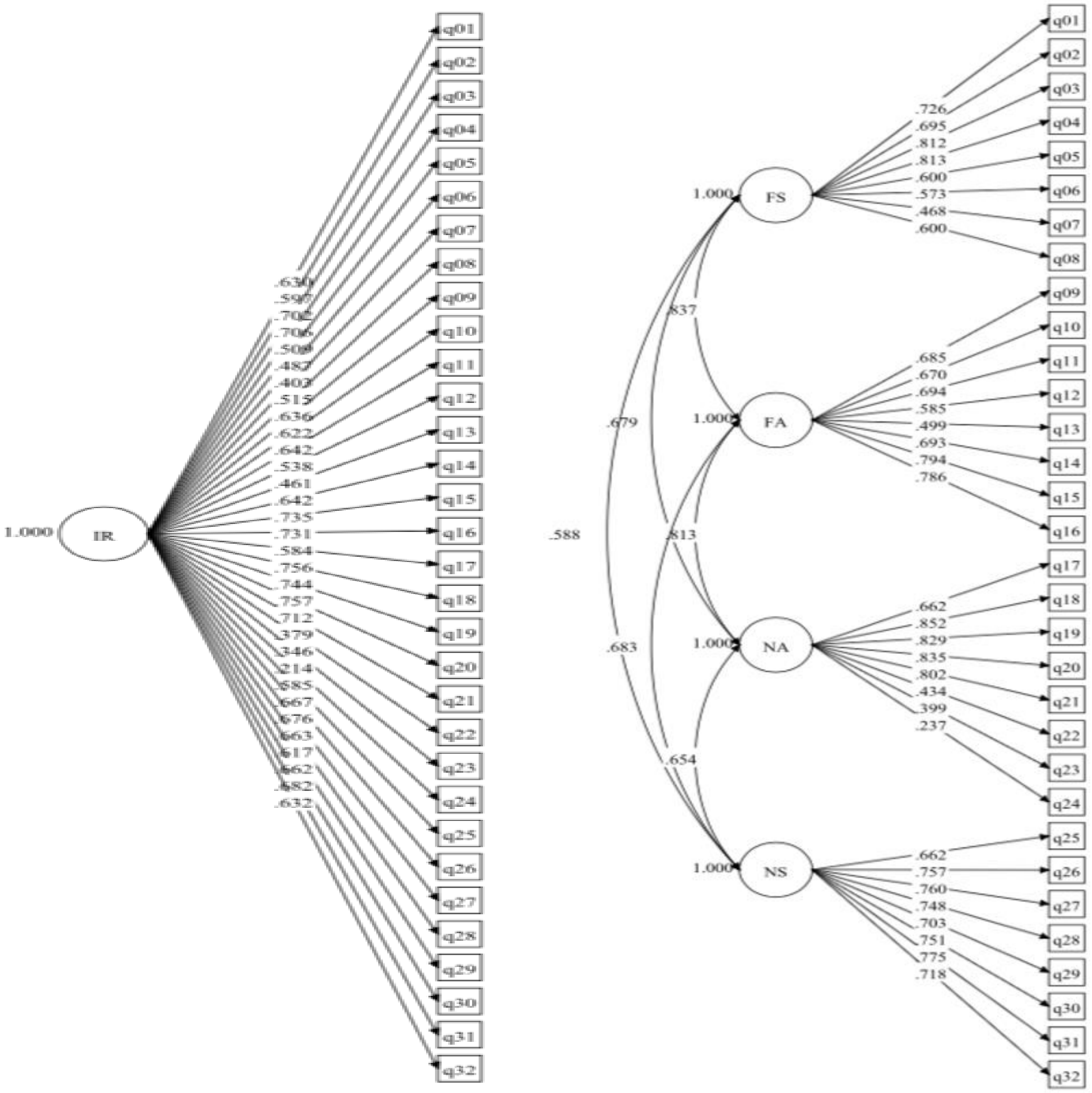
4.4.1 Test score reliability

Item internal consistency reliability estimates were calculated using R software with the psych package (Revelle, 2019). Cronbach's alpha values were calculated for the subtests and the whole test. The NS task achieved the highest alpha value of .81, and the lowest alpha coefficient, .69, was found for the NA. Cronbach's alpha was .88 for the whole test. No items were deleted to increase the reliability of the test. All the items appeared to be worth retaining. In general, the Cronbach's alpha in the adapted test indicated an acceptable level of internal consistency reliability.

4.4.2 Validity of the inductive reasoning test

4.4.2.1 Confirmatory factor analysis

The results of CFA showed that the model was a good fit to the values for the cut-off criteria: CFI = .908, TLI = .901, RMSEA = .050 CI (.046, .054) and WRMR = 1.50. In addition, 4-factor model was examined for the IR test with CFA, indicating that it was a well fit to the data, with: CFI = .972, TLI = .970, RMSEA = .028 CI (.023, .032) and WRMR = 1.052. As expected, the indicators all showed significant positive factor loadings with a significant level (see Figure 4.1). Overall, the model fit was at an acceptable level in both unidimensional and 4-dimensional structures, since the results for the indices were a good fit based on the cut-off standards.



a) Unidimensional construct b) 4-dimensional construct

Figure 4.1. Standardized factor covariances, item loadings from confirmatory factor analysis of (a) unidimensional and (b) 4-dimensional constructs for the IR test.

4.4.2.2 Rasch analysis

We analysed the data with item response modelling to investigate whether the IR test can measure students' IR proficiencies across school grades. The results of the Rasch analysis indicated a good fit model with the infit for single items (weighted mean squares, MNSQ) ranging from 0.84 to 1.22 (M = 0.98, SD = 0.09). The average item difficulty was fixed on 0 logits (SD = 1.0), representing the zero point of the scale, while the item difficulties ranged from −1.79 to 2.29. The quality of items on the test is good, while the discrimination values for most of the items were higher than 0.3 (for 31 out of 32 items) (Ebel & Frisbie, 1991). The relation between the average item difficulty of 0 logits and the average person proficiency of 1.08 logits on the MLE scale implied that students' proficiency in IR was higher than the average item difficulty. The final deviance in unidimensional model was 21,283.138 with 33 parameters estimate (AIC = 21,349.14, BIC = 21,377.03), while it is 21,390.177 with 42 parameters estimated in the four-dimensional model (AIC = 21,474.18, BIC = 21,509.7). The deviance in the four-dimensional model were higher than that in the unidimensional model, recommending that the unidimensional model had a better fit than the four-dimensional one. Overall, participants successfully completed 66.6% of the IR items (21.3 out of 32 items) on average. The psychometric properties of the items ordered from most difficult to least difficult are summarized in Table 4.2.

Table 4.2. The psychometric properties of the IR test in the Rasch model.

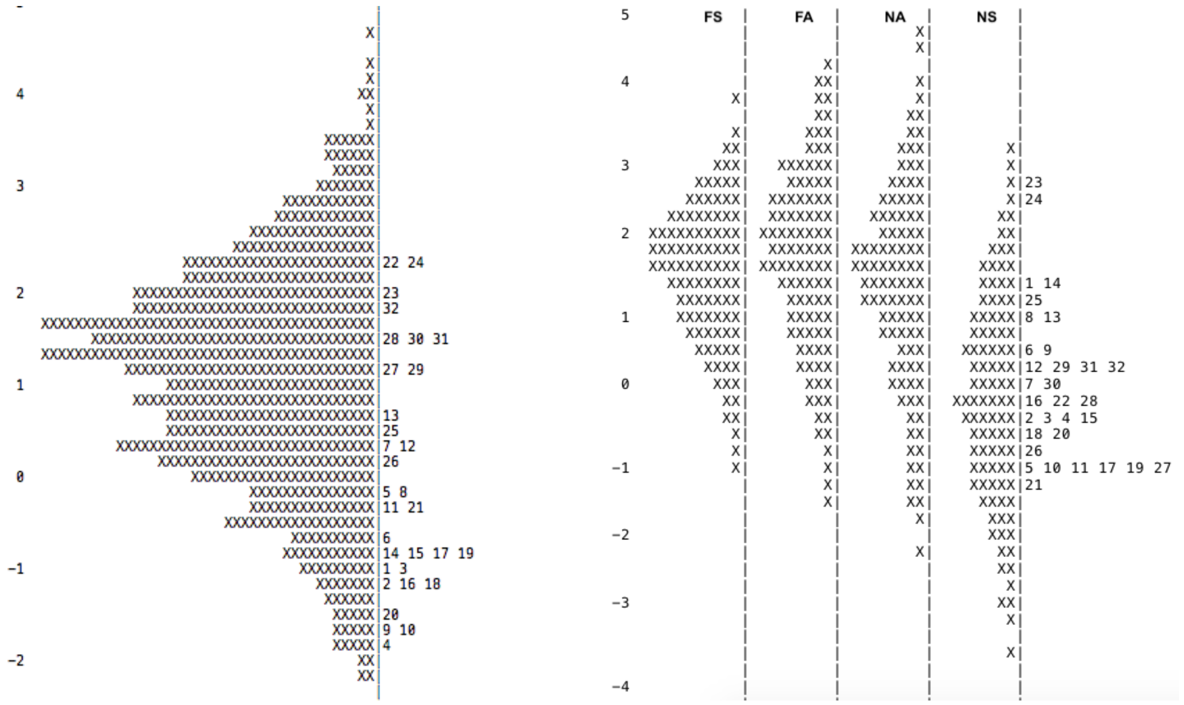
Item	Subtest	Correct answer (%)	Unidimensional model			4-dimensional model		
			Dis.	Difficulty	Infit	Dis.	Difficulty	Infit
1	FS	83.98	0.47	−1.06	0.96	0.07	1.40	1.31
2	FS	84.41	0.44	−1.10	0.98	0.46	−0.46	0.91
3	FS	83.71	0.51	−1.03	0.87	0.45	−0.50	0.89
4	FS	90.71	0.45	−1.79	0.89	0.52	−0.44	0.84
5	FS	71.71	0.42	−0.14	1.07	0.45	−1.17	0.86
6	FS	78.57	0.39	−0.61	1.07	0.42	0.40	0.98
7	FS	64.29	0.35	0.30	1.15	0.39	−0.04	0.98
8	FS	73.00	0.43	−0.22	1.04	0.36	0.81	1.05
9	FA	90.00	0.43	−1.69	0.94	0.44	0.36	1.11
10	FA	89.14	0.42	−1.58	0.94	0.43	−1.15	0.96
11	FA	75.29	0.52	−0.37	0.98	0.43	−1.06	0.96
12	FA	64.14	0.46	0.31	1.05	0.52	0.20	0.96
13	FA	57.54	0.41	0.70	1.11	0.46	0.92	1.03
14	FA	82.29	0.49	−0.89	0.96	0.41	1.30	1.13
15	FA	80.86	0.56	−0.77	0.87	0.49	−0.35	0.97
16	FA	85.71	0.53	−1.20	0.86	0.57	−0.24	0.87
17	NA	81.86	0.44	−0.86	0.98	0.53	−1.13	0.95
18	NA	84.86	0.54	−1.12	0.84	0.44	−0.75	0.99
19	NA	81.29	0.53	−0.81	0.92	0.54	−1.04	0.86
20	NA	88.14	0.48	−1.46	0.93	0.54	−0.70	0.91
21	NA	73.86	0.55	−0.28	0.92	0.48	−1.39	0.86
22	NA	27.71	0.33	2.28	1.14	0.56	−0.10	0.89
23	NA	32.43	0.32	2.00	1.14	0.33	2.71	1.14
24	NA	27.57	0.22	2.29	1.22	0.31	2.40	1.14
25	NS	61.00	0.50	0.48	0.99	0.22	1.19	1.34
26	NS	66.14	0.55	0.19	0.91	0.49	−0.80	1.04
27	NS	49.86	0.56	1.07	0.87	0.55	−1.11	0.94
28	NS	42.29	0.53	1.46	0.93	0.56	−0.15	0.93
29	NS	48.43	0.51	1.14	0.99	0.52	0.29	0.92
30	NS	42.43	0.52	1.45	0.94	0.49	−0.07	0.99

31	NS	41.14	0.52	1.52	0.93	0.51	0.29	0.93
32	NS	36.14	0.48	1.79	0.98	0.49	0.35	0.95

Note. FS: figure series completion; FS: figure analogies; NA: number analogies; NS: number series completion, Dis.: Discrimination.

The Wright person-item map depicted in Figure 4.2 shows the level of ability of students on the left-hand side and the difficulty of the items on the right. This item-person map describes the distribution of item difficulty estimates and person ability estimates on one scale, making it possible to compare items and persons directly within the context of the study and judge if the difficulty of the items was appropriate for the actual participants. If a student is located at the same level as an item, this suggests that the student has a 50% chance of responding to that item correctly (p = .5). If the student is plotted higher than the item, then the chances of success increase (p > .5) and will continue to increase as the distance between the person and the item widens. Similarly, if the item is placed higher than the student, the student will not be likely to succeed; indeed, the chances of success decrease if the distance between the item and the person increases (Griffin, 2010). The distance of the item from the top of the ruler correlates to its difficulty relative to the other items. In this study, the most difficult items (22 and 24) are shown at the top on the right of the y-axis. Item 4 is the easiest since it stands at the bottom of the right-hand side of the map.

A general pattern of item difficulty can clearly be seen within the abilities of the test-taking population in Figure 4.2. The item-person map summarizing the test performance of the sample confirms that many of the test-takers are quite beyond the higher limit of the test. Item 4 is too easy and falls outside the abilities of the students. According to Griffin (2010), good items have to cover all the areas on the ruler when measuring the ability spectrum of all students. All in all, despite the existence of a little misfit for items, this test appears to have performed well as a measurement instrument. The test is appropriate for the assessment of IR for the study sample. The items selected for this study did cover most of the persons on the scale, so it seems that the test is well-targeted for this group of test-takers.



a) Unidimensional mode

(Each ‘X’ represents 1.1 cases).

b) 4-dimensional model

(Each ‘X’ represents 6.8 cases).

Figure 4.2. Wright map of persons and items for the (a) unidimensional and (b) 4-dimensional models.

Note. FS: figure series completion; FS: figure analogies; NA: number analogies; NS: number series completion

4.4.3 Differences in students’ performance among grade cohorts

This study also aims to investigate the effects of grade levels on students’ performance on the IR test. We composed pirate plots to visualize the distribution of achievements among students in the study. A pirate plot can easily show raw data, descriptive statistics and inferential statistics in one plot. These pirate plots for each group included a box plot with the mean, 95% shaded highest density intervals, the jittered individual data points and symmetric kernel densities. They provided more information than regular bar plots and box plots (Phillips, 2016, 2017). Figure 4.3 depicts students’ latent abilities on the IR test across grade cohorts. Because the intervals among the sample did not overlap, we can confidently conclude that students’ mean abilities on the test showed a remarkable improvement across cohorts. In the 5th-grade group, there was a combination of a number of students who received a very low score and the smallest number of students who achieved a high score. On average, the 5th graders yielded a score of 0.24 logits. The mean scores in the 7th- and 9th-grade groups were around 0.92, but their distribution of scores showed a different pattern. The 7th-grade cohort had both the highest- and lowest-performing participants. It appeared that there was an equivalent proportion of students who achieved higher mean scores and lower scores in this group. In the 9th-grade group, the number of students achieving a higher mean score (M = 1.25) was larger than that of those who received a lower mean score. Among the 11th graders, the average mean remained at about 1.86, but there were still some students receiving under 0. The distribution is spread out and has its highest density because of the higher rate of students of high ability. In the study, we plotted the smooth curve to present the trend in students' performance on the IR test across grade groups (with the ggplot2 package, see Wickham, 2016). Figure 4.3 illustrates the change of students' IR from the 5th to 11th grades. In general, students' achievement on the IR test grew gradually throughout the grade levels. The strongest growth occurred from the 5th to 7th grades with Cohen's d effect size of 0.54, and the trend slowed down at the 7th to 9th grades (Cohen's d = 0.26). The growth rate tended to speed up again from the 9th to 11th grades (Cohen's d = 0.50).

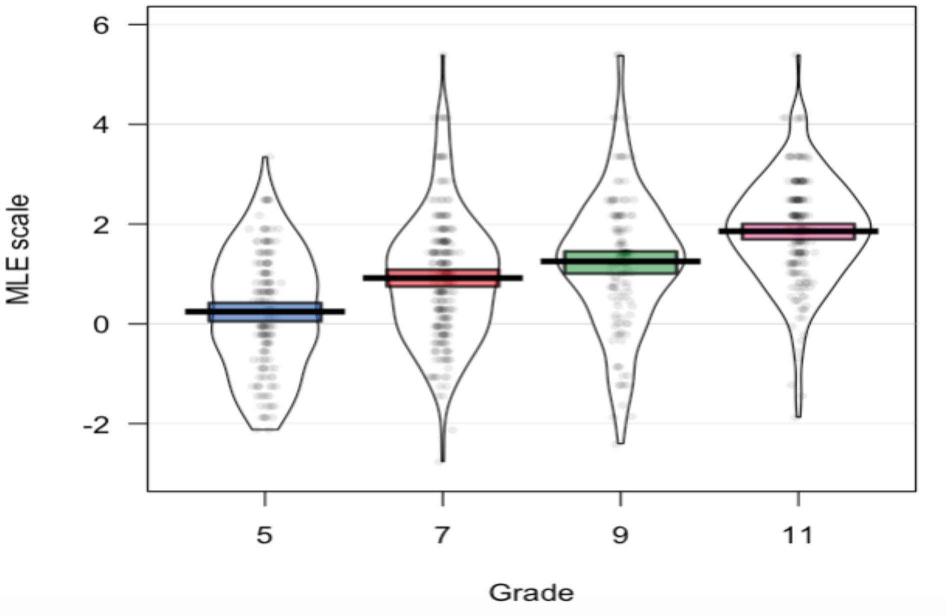


Figure 4.3. Differences in the students' performance on the IR test regarding grade cohorts.

As regards children's reasoning proficiency on each subtest, students achieved different proficiency on different kinds of subtests. Students' latent abilities were lowest on the NS subtest (M = 0.3, SD =1.9) and highest on the FA subtest (M = 2.0, SD = 1.6), followed by the FS (M = 1.5, SD = 1.5) and the NA (M = 0.5, SD = 1.6) subtests.

Figure 4.4 depicts the smooth curves in four different subtests that students exhibited throughout grade cohorts. Overall, there was different achievement among students on the different tasks across school grades. It is easy to recognize that students performed better on the subtests constructed with figures, such as the FS and FA subtests. Participants showed the highest proficiency levels on the FA task, followed by the FS task throughout the sample cohorts. The results of test-takers on the NA task showed a different pattern, since it fluctuated from the 7th to the 9th grades. The 7th-grade students achieved very high, even higher than the 9th graders did, on the NA subtest. In contrast, there was a linear increase in the students' performance on the NS task from just under −1.0 in the 5th grade to around 1.0 in the 11th grade. Student achievement on the NS seemed to strongly depend on school grade levels.

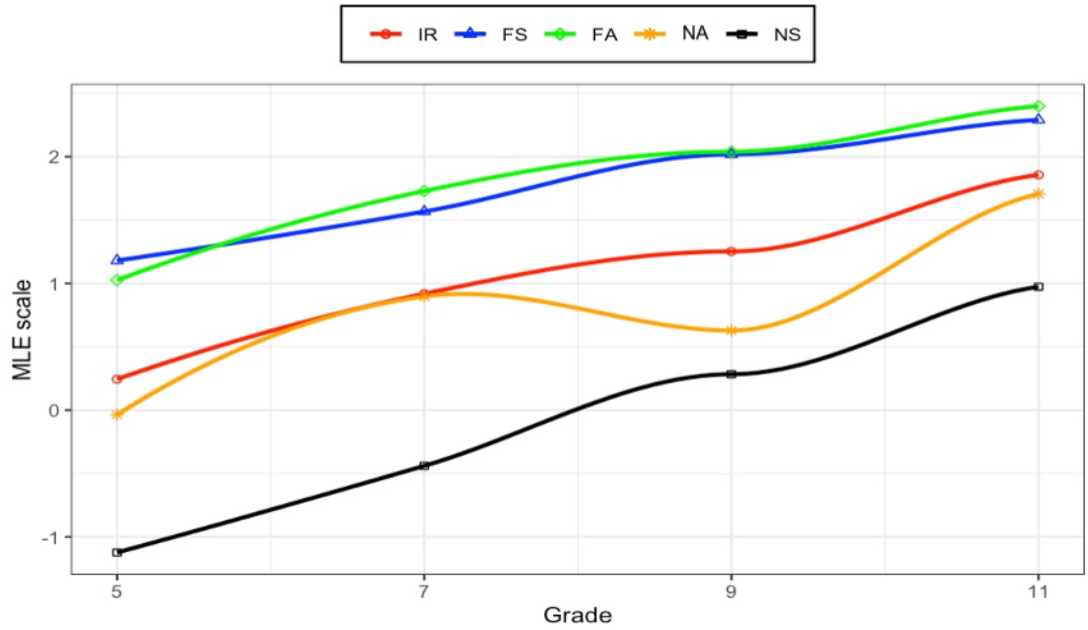


Figure 4.4. Development of reasoning across grade levels on the test and subtests.

Note. IR: inductive reasoning test; FS: figure series completion task; FS: figure analogies task; NA: number analogies task; NS: number series completion task.

Furthermore, the ANOVA analysis was conducted to compare the effect of school grade groups on students' proficiency on the subtests and the test. The results revealed significant disparities among the grade cohorts on both the subtests and entire test: FS task [F(3, 697) = 20.78, p < .01], FA task [F(3, 697) = 25.61, p <. 01], NA task [F(3, 697) = 36.83, p < .01], NS task [F(3, 697) = 44.63, p <. 01] and entire test [F (3, 697) = 51.76, p < .01]. To identify whether there were significant differences in pairs of grades on the test and subtests, Tukey's HSD analysis was performed, and the results were summarised in Table 4.3.

A significant difference was found for all pairs of grade levels on the IR test with effect size from small to medium, except for a pair in the 7th and 9th grades. As regards the subtests, the students in the upper grades performed significantly better than ones in the lower grades on the NS subtest. This also happened on the NA subtest, except for a 7th- and 9th-grade pair, where the 7th graders achieved a higher mean score than the 9th graders did. The general trend in students' performance on the FS subtest and the FA subtest was similar across grade levels.

However, the school grade difference was not found on the FS task in the pairs of 11th- and 9th-grades. Overall, the results indicated that the students from the higher grades tended to perform significantly better on the numerical tasks than the students from younger cohorts did, but students earned a higher score on the figural tasks than they did on the numerical tasks.

Table 4.3. Tukey HSD’s multiple comparisons.

Subtest	7th & 5th		9th & 7th		11th & 9th	
	ΔM	Cohens’ d	ΔM	Cohens’ d	ΔM	Cohens’ d
FS	0.39	0.26*	0.45	0.32**	0.27	0.20
FA	0.70	0.44***	0.31	0.21*	0.35	0.26*
NA	0.93	0.57***	−0.27	0.17	1.08	0.73***
NS	0.68	0.37***	0.72	0.38***	0.69	0.40***
IR	0.66	0.54***	0.27	0.26**	0.63	0.50***

Note. ΔM: Mean difference.
*p<.05, **p< .001, ***p< .001.

4.4.4 Gender differences in inductive reasoning

The general trend was that males and females showed an equivalent proficiency in each cohort (Figure 4.5). The mean scores stayed at around 0.2 and 1.0 for the 5th grade and 7th grade, respectively. In the 9th grade and 11th grade, the average score among males and females was at a similar level, but these distributions displayed different patterns. In comparison with girls in the 9th grade, the larger part of the density of the male sample shifted toward both the minimum and maximum scores. This demonstrated that the standard deviation was larger in the male cohort. In contrast, male students in the 11th grade had a smaller standard deviation than the females.

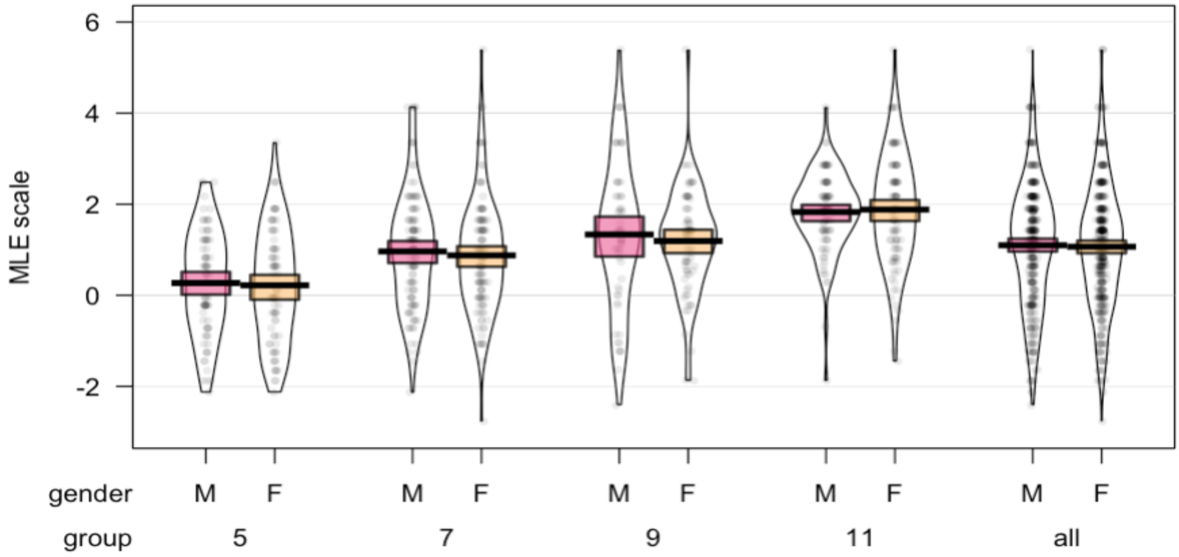


Figure 4.5. Comparing performance among males and females.

We applied the *t*-test to compare abilities between males and females. Table 4.4 presented the results of the *t*-test in the four cohorts and whole sample. No significant difference was found on the IR test between boys and girls in the whole sample and in individual cohorts (p > .05). This also suggested that students' abilities in each cohort did not differ significantly on the IR test regarding gender. In other words, it was estimated that males and females had an equivalent ability level in IR.

Table 4.4. The *t*-test to compare inductive reasoning test results between males and females.

	Male		Female				
Grade	N	Mean (SD)	N	Mean (SD)	t	p	Cohens' d
5	78	0.27 (1.14)	79	0.22 (1.24)	0.27	.791	0.04
7	107	0.96 (1.27)	115	0.88 (1.26)	0.51	.610	0.07
9	57	1.33 (1.64)	78	1.19 (1.13)	0.57	.572	0.10
11	90	1.83 (0.88)	97	1.88 (1.23)	−0.34	.731	0.05
All	332	1.09 (1.35)	369	1.07 (1.35)	0.31	.754	0.02

Moreover, no gender difference was indicated for the single subtests. As provided in Table 4.5, all p-values were larger than .05. In other words, gender had no significant impact on performance among students on the subtests. However, it is surprising that male students achieved a slightly higher mean score than female students on tasks with figural material, but this was not found on tasks with numerical material, where female students achieved a higher mean score than male students. Standard deviations for the males' abilities appeared somewhat higher than those of the females.

Table 4.5. The *t*-test to compare latent abilities between males and females on each subtest.

Subtest	Male	Female			
	Mean (SD)	*Mean (SD)*	t	p	Cohens' d
Figure series completion	1.82 (1.43)	1.71 (1.48)	1.05	.292	0.08
Figure analogies	1.84 (1.64)	1.78 (1.48)	0.49	.621	0.03
Number analogies	0.87 (1.68)	0.84 (1.67)	0.20	.840	0.02
Number series completion	−0.08 (1.88)	−0.07 (2.01)	−0.11	.915	0.01

4.4.5 Predicting students' inductive reasoning performance

Stepwise multiple regression was used to explore how the latent factors predicted IR ability in individuals. The results of a stepwise multiple regression analysis were presented in Table 4.6. School grade group, school performance in the previous semester and parents' educational level significantly explained 32.0% of the variance on the IR test, F (680) = 79.0, p < .001. School performance grade in the previous semester was the best predictor with standardized beta regression coefficients of .65 each, followed by school grade groups with .33 each. Parents' educational level significantly affected IR ability in children. Specifically, the effect of educational level among students' fathers on IR outweighed that of their mothers' education in the study sample.

Table 4.6. Hierarchical regression summary for prediction of inductive reasoning (N = 701), $R^2 = .32$, adjusted $R^2 = .31$.

Model	M1	M2	M3	M4
Predictor	β	β	β	β
School grade	.26***	.31***	.33***	.33***
School performance		.76***	.69***	.65***
Mother's education			.17***	.11*
Father's education				.14**

Note. $\Delta R^2 = .18$ for M1, $\Delta R^2 = .11$ for M2, $\Delta R^2 = .02$ for M3, $\Delta R^2 = .01$ for M4;

* p< .05, ** p< .01, *** p< .001

4.4.6 Validity and reliability of the SMTSL questionnaire

As suggested by Dermitzaki et al. (2013), we employed the fit testing for the SMTSL questionnaire with a bifactor model to the empirical data. The bifactor model involves four independent motivational factors, including self-efficacy (SE), active learning strategies (AL), science learning value (LV), achievement goals, (AG) learning environment stimulation (LE), and science motivation in general (SM). The results showed that the model fit is acceptable but not excellent, with cut-off criteria ($\chi^2(139) = 763.3116$, p < .001, CFI > .901, RMSEA <. 080, WRMR = 1.627). All the indicators provided significant positive factor loadings as depicted in Figure 4.6.

We used the internal consistency index of McDonald's omega (ω) in R package psych (Revelle, 2019) for reliability estimates, since it is less biased than Cronbach's alpha in this case (Dunn et al., 2014). Internal consistency reliability was generally adequate, with an omega of .87 for the whole items. It was also an acceptable level for the single subscales: SE (ω = .69), AL (ω = .80), LV (ω = .56), AG (ω= .74) and LE (ω = .74).

Furthermore, DIF analysis was used to test measurement invariance with respect to gender with the R lordif package (Choi et al., 2011). The McFadden's R^2 is set at 0.01 as a criterion for rejecting the null hypothesis of no DIF. All pseudo R^2 values were below 0.01, and p-values of the goodness-of-fit statistics above 0.05, indicating no item were flagged as DIF item regarding gender. Overall, these results suggest that the models were well supported with the empirical data. The results are consistent with the characterisation of motivational factors proposed in the literature (Cavas, 2011; Chan & Norlizah, 2018; Dermitzaki et al., 2013; Shaakumeni & Csapó, 2018; Tuan et al., 2005).

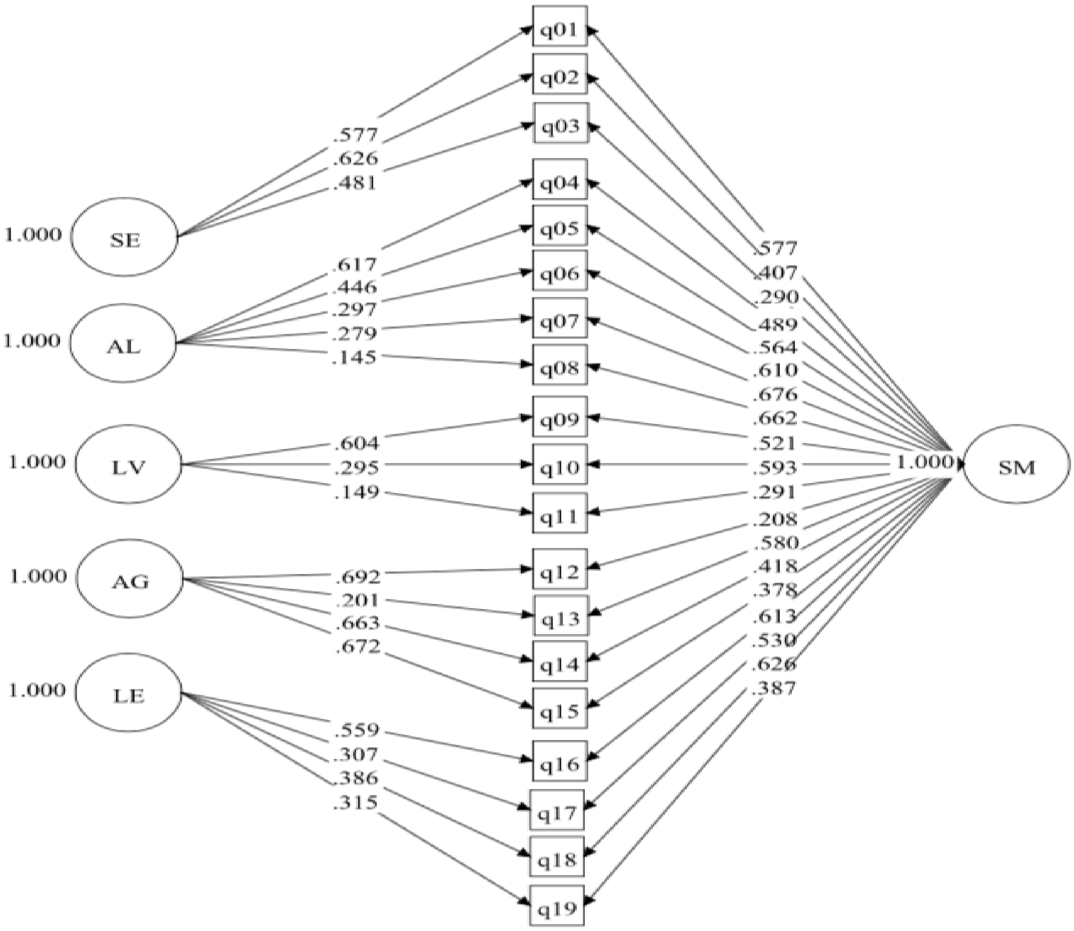


Figure 4.6. Bifactor model of the SMTSL questionnaire.

4.4.7 Differences in science motivation among students between grade levels

Figure 4.7 illustrates the differences of students' motivation in each subscale and science motivation in general between grade cohorts. We plotted the smooth curves (the dotted line) to visualize the patterns of students' science motivation across grade levels (see more at Wickham, 2016). Generally, students' science motivation gradually decreased across grade levels.

As regards separate subscales, the students seemed to get the lowest scores on the LE scale across the grade groups. The younger grades (5th and 7th) achieved the highest score on AL scale. The scores for this scale fell sharply from 4.26 in the 5th grade to under 3.81 in the 11th grade. There was a noticeable reduction in the scores for SE from above 4.23 in the 5th grade to just 3.46 in the 11th grade. Although the scores fluctuated toward the AG and LV scales, the older children scored a little lower than the younger ones. The younger groups (5th and 7th graders) seemed to receive higher self-efficacy and motivation scores in learning science, while the older groups tended to be more concerned about the values and aims of learning science.

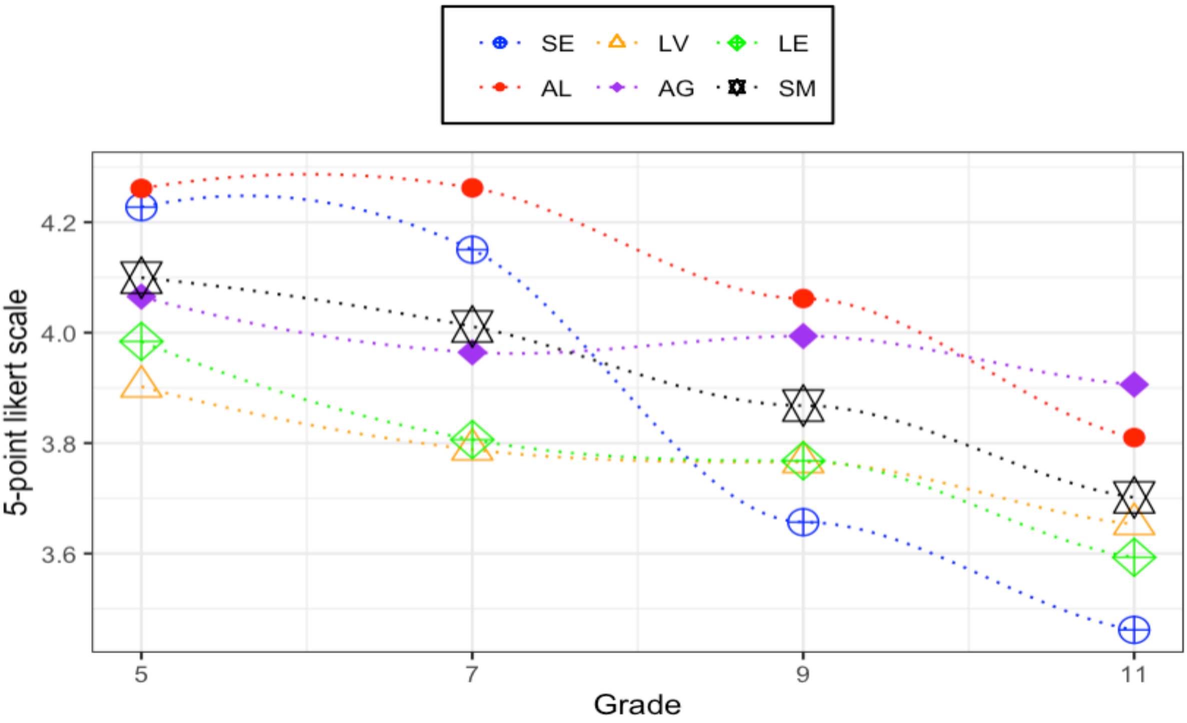


Figure 4.7. Changes of science motivation on the subscales across grade cohorts.

Note. SE: self-efficacy; AL: active learning strategies; LV: science learning value; AG: achievement goals; LE: learning environment stimulation; SM: science motivation in general.

Overall, the students showed a high positive perceived motivation towards learning science. The highest score was recorded for the AL (M = 4.10, SD = 0.54), followed by AG (M = 3.98, SD = 0.56) and the SE subscale (M = 3.89, SD = 0.79). The LV had an average score of 3.77 (SD = 0.65), while LE had the lowest score (M = 3.78, SD = 0.65).

In addition, we manipulated the ANOVA analysis to explore the influence of grade levels on the individual subscales and whole questionnaire. Except in the AG subscale, a significant difference was found between the grade cohorts in most subscales: SE [F(3) = 46.81, p<.001], AL [F(3) = 32.65, p < .001], LV [F(3) = 4.27, p = .005], LE [F(3) = 10.99, p < .001] and SM in general [F(3) = 5.40, p < .001]. Tukey's HSD analysis was used to identify which grades showed significant differences statistically. Table 4.7 provides the results of Tukey HSD's multiple comparisons between grades. Despite the older students being lower than that of the younger ones in general science motivation, the value and achievement goal of children in science subjects seemed unchanged among students in different grades.

Table 4.7. Tukey HSD's multiple comparisons.

Subscale	7th & 5th		9th & 7th		11th & 9th	
	ΔM	Cohens' d	ΔM	Cohens' d	ΔM	Cohens' d
SE	−0.08	0.11	−0.49	0.70***	−0.20	0.25*
AL	0.00	0.00	−0.20	0.39**	−0.25	0.50***
LV	−0.11	0.18	−0.02	0.04	−0.11	0.17
LE	−0.18	0.29*	−0.04	0.06	−0.17	0.27*
SM	−0.09	0.23	−0.14	0.35**	−0.17	0.38**

Note. SE: self-efficacy; AL: active learning strategies; LV: science learning value; LE: learning environment stimulation; SM: science motivation in general.

*p<.05, **p<.01, ***p<.001.

4.4.8 Gender difference in students' motivation toward science learning

Figure 4.8 illustrates the performance of male and female students on each motivation subscale and general scale. It seems that girls had higher scores than boys on most of the subscales of motivation, except on the LE subscale. Furthermore, the *t*-test showed that a significant disparity was only found on the achievement goals subscale, with the girls achieving higher scores (M = 4.05, SD = 0.53) than the boys (M = 3.89, SD = 0.61), t(760.5) = 3.50, p < .001. Girls did not differ significantly from boys in most of the subscales and general science motivation in this study.

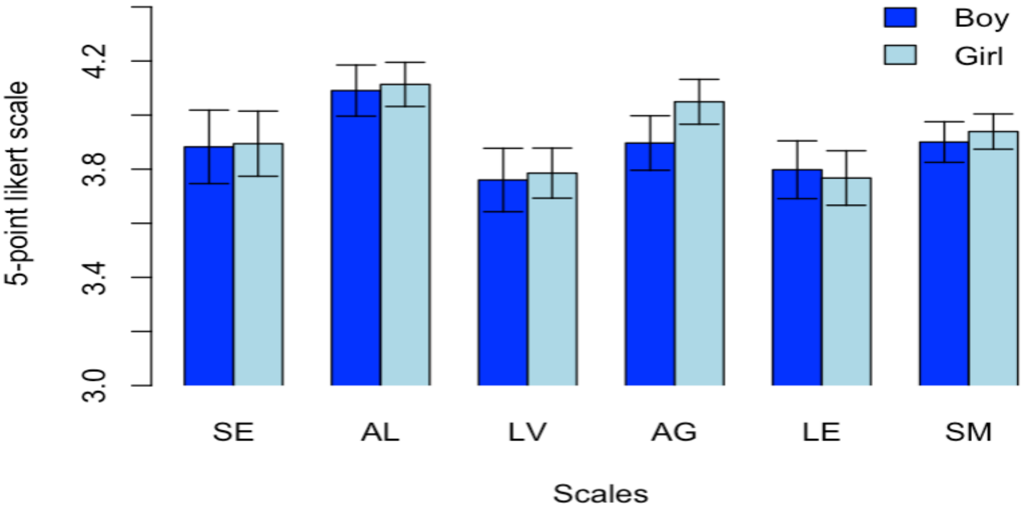


Figure 4.8. Comparison of science motivation among males and females on each subscale. Note. SE: self-efficacy; AL: active learning strategies; SL: science learning value; AG: achievement goals; LE: learning environment stimulation; SM: General science motivation.

Reflecting differences for each subscale among the grade cohorts, Figure 4.9 demonstrates performance patterns among males and females across the grade levels. A fluctuation in motivation level among boys and girls can be observed on the individual subscales, but the general tendency was that girls reported higher motivation than boys, especially on the AG and LE subscales. The males performed lower than the females on all the subscales, except in the 11th grade, where the boys achieved higher scores than the girls.

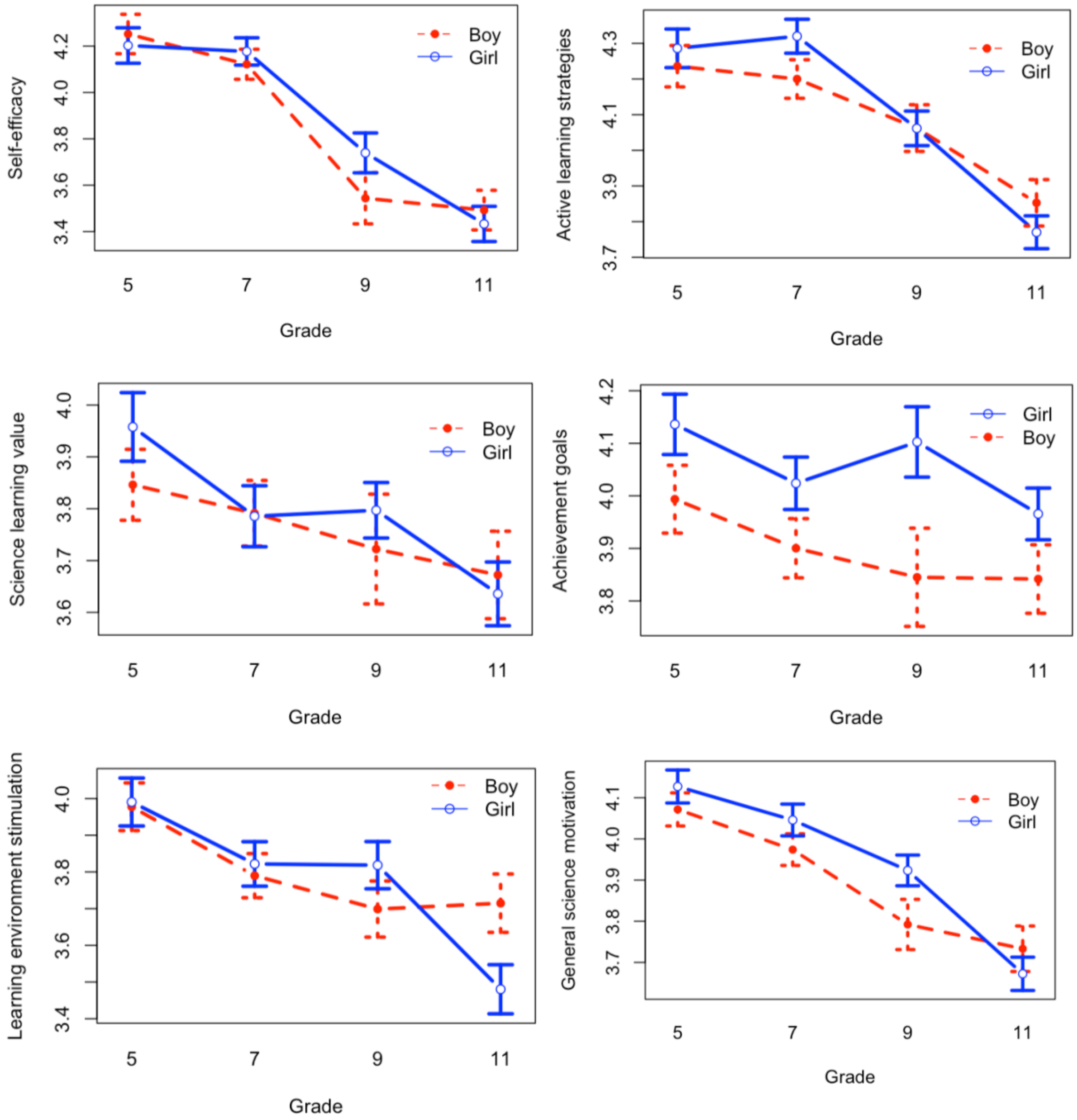


Figure 4.9. Performance of males and females on each subscale and whole questionnaire among grade cohorts.

Additionally, we continued to examine the gender difference with the t-test within each subscale across school grade cohorts. No significant gender difference was found between males and females on any subscale across grade levels ($p > .05$), except in the 9th grade where girls reported higher scores on achievement goal [$t(107.6) = 2.24$, $p = .027$] and in the 11th grade with a higher point for boys in LE subscale [$t(177.2) = 2.25$, $p = .025$]. It may be concluded that students' motivation toward science learning is not dependent on gender in the current study.

4.4.9 Relationship between inductive reasoning and science motivation

Pearson's product-moment coefficient was calculated to investigate the relationship between IR and SM. Table 4.8 summarizes the correlation between these variables on the individual subscales and on science motivation scale in general. Overall, the relationship between IR and SM was not strong. There was a positive correlation between the two variables across the grades, except in the 11th grade. IR was found to be positively correlated with most of the subscales in the 5th, 7th and 9th grades. In the 11th-grade cohort, IR was negatively related to both the subscale and general scale. It appears that the self-efficacy subscale has a clearer link to IR than to other ones, while AG and LE subscales showed no significant relation to IR.

Table 4.8. Correlation (Pearson) between IR and SM.

Scale	Grade			
	5th	7th	9th	11th
Self-efficacy	.15	.22***	.29***	−.06
Active learning strategy	.13	.12***	.08	−.02
Science learning value	.01	.21**	.09	−.08
Achievement goals	.10	.04	.05	.02
Learning environment stimulation	.09	.00	−.10	−.14
General science motivation	.16*	1.2**	.14	−.07

Note. *p<.05, **p<.01, ***p<.001.

4.4.10 Main factors predicting science motivation

STEM achievement was referred through manipulating an average of four school subjects (math, physics, chemistry and biology) in the last semester. We included inductive reasoning (IR), student age (AG, a representation of grade level), gender (GE), school difference (SC), STEM achievement (STEM), mother's education level (ME), father's education level (FE) and parental involvement (PI) as the explanatory variables in our exploration.

We employed the BMA analysis (see Section 3.3.5) to explore the proposed model with linear regression models in R package BMA (Raftery et al., 2020). The best models were recommended based on the BIC values for each manipulation. For instance, when exploring latent predictors for the SE subscale, the model, which included IR, AG, STEM and PA, was suggested, since it had the posterior model probability (21.7%) and the lowest BIC index (−678.14) and explained around 21.0% of explained variance. Similarly, we employed a BMA analysis for other scales and found that the most frequent variables in these models were IR, AG, SA, ME and PI, while SC and FE were absent in most of the models.

Furthermore, we implemented a path analysis in SEM approach to test for direct and indirect effects. After investigating several models, the one with five main explanatory variables of motivation has proven the best. Finally, our results offered a general model illustrated simply in Figure 4.10, which presents significant relations between latent variables. Results also show that the models fit well the present data on all the scales (Table 4.9).

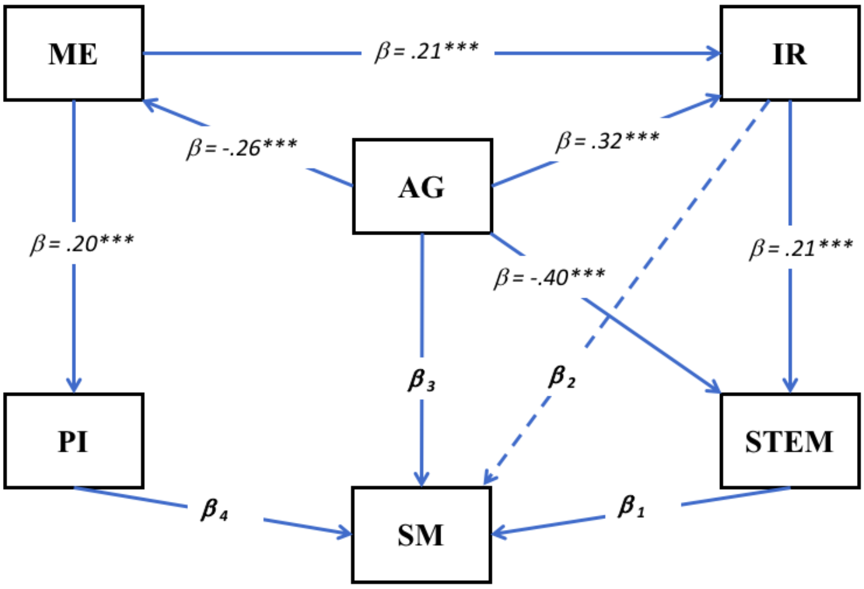


Figure 4.10. The general model of relations between explanatory variables predicting SM. Note. ME: mother's education level; IR: inductive reasoning; AG: student age; PA: parental involvement; STEM: STEM achievement; SM: science motivation in general. *β*: regression coefficient.

As regards the role of IR in students' SM, it seems that IR showed a direct and indirect effect on motivation, in which STEM achievement was expected as a mediator. The mediated relationship was examined with the Sobel test, and it was determined that the mediating effect was significant on the subscales for SE (Z = 2.87, p = .004), LV (Z = 2.40, p = .017) and SM in general (Z = 2.23, p = .026). IR has a positive relation with all the SM subscales, except the LE subscale, where students who achieved a higher score in IR reported lower motivation. Mother's education level directly affects student motivation, but a relationship between the two variables is established through parental involvement. Age has a negatively significant relationship with SA, SE and AL factors, but it is positively linked to IR.

Table 4.9. The model fits and regression coefficients in predicting science motivation.

Scale	χ^2(5)	p	CFI	RMSEA	SRMR	β_1	β_2	β_3	β_4
SE	18.49	.002	.952	.079	.046	.17***	.10*	−.30***	.19***
AL	16.18	.006	.962	.072	.042	.09*	.04	−.23***	.36***
LV	17.65	.003	.941	.075	.042	.14**	.06	.03	.20***
AG	16.38	.006	.942	.072	.042	−.06	.02	−.03	.20***
LE	15.94	.007	.946	.071	.041	.08	−.11*	−.01	.19***
SM	15.94	.007	.960	.071	.042	.12**	.03	−.19***	.32***

Note. SE: self-efficacy; AL: active learning strategies; LV: science learning value; LE: learning environment stimulation; SM: science motivation in general.

*p<.05, **p<.01, ***p<.001.

4.5 Discussion and conclusion

The results indicated that the adapted version of the IR test demonstrated as a reliable instrument for measuring reasoning in school age groups in the 5th, 7th, 9th and 11th grades. This suggests that the IR test may be used in a broad age range (hypothesis H1.1 is confirmed). Similarly to previous paper-based IR tests, for example, in a cross-sectional study, Csapó (1997) used the earlier version of an IR test from the 3rd grade to the 11th grade. The average person proficiency in IR was higher than the average item difficult, and participants completed most of the items correctly. The item-person map shows that the construction of the IR test needs to be improved to measure students' performance at the higher end of the proficiency continuum. The present study demonstrates that the test, comprising figural and numerical material, was a reliable tool to assess IR. Students tended to achieve a better performance with high scores on items with figural material than items with numerical material. These findings correspond with that of the study conducted by Roberts et al. (2000).

Analogously, CFA and PCM analysis also confirmed hypothesis H1.1 in which the instrument provides an adequate fit to the data and is suited to measuring SM in the Vietnamese context. The results are consistent with past studies using the same questionnaire in other contexts: Namibia (Shaakumeni & Csapó, 2018), Malaysia (Chan & Norlizah, 2018), Turkey (Cavas, 2011), and Greece (Dermitzaki et al., 2013). The results revealed that Vietnamese students showed high motivation in learning science and positive attitudes toward science.

Students' performance on the IR test increased across school grade cohorts. The findings are seemingly consistent with the findings of Csapó (1997), Molnár and Csapó (2011), Molnár et al. (2013), Díaz-Morales and Escribano (2013) and Csapó et al. (2019). A series of these measurements concluded that the development of IR was significant across grade levels. This is relatively responding to the hypothesis H1.2. The fastest development happened between the age group of 12-14 years or during the middle school educational level, but the acceleration of improvement seemed to slow down after 14 years. This suggests that it is the most effective time to enhance IR. Thus, teachers should be aware of this opportunity and actively search for means to boost students' reasoning skills during those years through their school subjects.

As regards gender, no statistically significant difference was found in IR and SM. Boys and girls showed no significant difference in their performance in single grade cohorts and even for the whole study sample. For the IR test, the results were in line with previous studies in Namibia (Kambeyo, 2018) and Spain (Díaz-Morales & Escribano, 2013), although males achieved a slightly higher mean score compared to females, in agreement with the general conclusion of the meta-analysis performed by Waschl and Burns (2020). However, although there was no significant difference between the boys and girls in science motivation in general, a slight difference was observed in favour of the females. These findings are in keeping with a study by Józsa (2014) in Hungary. Nevertheless, the studies in Turkey (Cavas, 2011) and Malaysia (Chan & Norlizah, 2018) found that females reported significantly higher scores than males. The findings partly endorsed the hypothesis H1.3.

These findings have contributed to an increased understanding of latent factors predicting the IR ability of children. Learning in the disciplines mirrored in the school performance in the previous semester prominently enhanced IR. Along with the age group factor, parents' education level impacts appreciably on individuals in terms of IR. Students' IR proficiency is especially affected by the educational levels of their fathers. The hypothesis H1.4 is confirmed. However, the results are inconsistent with the outcome of the previous study, in which mothers' educational level had a stronger impact than that of the fathers (Csapó, 2001) or parents' educational level had no influence on students' achievement on the IR test at all (Kambeyo, 2018). These differences in the impact of parents' education may be attributed to cultural differences and to the different impact of the socio-economic status of the families, as is routinely reported in the PISA surveys. In the PISA 2015 assessment, Vietnam was among those countries where the impact of students' socio-economic status on performance was the lowest (OECD, 2016).

As regards the change of motivation in the different grade cohorts, the results were mostly consistent with the literature (Bouffard et al., 2001; Dorfman & Fortus, 2019; Józsa et al., 2017), indicating that student motivation tended to decrease gradually during secondary school education. The current study also found that students’ science motivation dropped from primary to upper secondary education. In other words, the hypothesis H1.2 is confirmed. When reaching upper secondary education, students tend to be less motivated than those in the early secondary grades on most of the subscales. Nonetheless, the students showed an unchanged motivation toward values and purpose of learning science, but their positive perspectives on their own abilities and learning environment declined grade by grade from the primary to the secondary education levels. This may be due to their being more concerned about the matriculation examination as a general expectation of their parents after they enter high school (Du, 2015). The findings play an important role in identifying the priority factors in enhancing science motivation in school practice. They signal a potentially alarming picture of change in students’ motivation and call for more facilitators of science inspiration and more concern with regard to enriching school environment factors in high schools in Vietnam.

Furthermore, the results of the BMA analysis contributed to an increased understanding of latent factors and models to predict students’ science motivation. Despite a positive correlation between IR and SM in most of the grade cohorts, the former does not directly affect the latter. The results support the hypothesis H1.5 and hypothesis H1.6. The analyses confirmed that students’ motivation in learning science was tied to learning science subjects in schools, and such students’ achievement in learning science disciplines in the previous semester meaningfully affected their motivation. The outcomes agree with those of the studies in Taiwan (Tuan et al., 2005) and Malaysia (Chan & Norlizah, 2018). Interestingly, parental involvement was observed as one of the core predictors of SM, while parents’ education levels seemed not to be linked to students’ motivation in schools. Ample parental support and engagement as well as interest in school activities play a key role in children’s motivation in learning and school (OECD, 2017b). This is due to the typical culture of the traditional family in Vietnam, where parents (especially the mother) are greatly interested in the performance and activities of their children in school (Hoang et al., 2014; Phan, 2004). Additionally, the model provides an interesting relation between student age and mother’s education level, in which the mothers of younger student groups seem to have higher education levels than those of older ones. This reflects the current situation of improvement in females’ education level in Vietnam in recent years.

However, even though the students reported high motivation toward learning science, this may cause potential risks to external pressure on students from their parents because the PISA results showed that those who experience high motivation tend to feel anxious about a test, even if they were well-prepared (OECD, 2017b). Hence, both schools and families should be concerned about how to inspire students’ motivation to learn and reduce excessive fear of failure. Teachers, school leaders and school psychologists should be aware of these impacts to create a more supportive learning environment.

CHAPTER 5. RELATIONSHIP BETWEEN INDUCTIVE REASONING, SCIENTIFIC REASONING AND SCIENCE MOTIVATION, AND THEIR ROLES IN PREDICTING STEM PERFORMANCE

5.1 Introduction

Improving students' general cognitive abilities is one of the main declared goals of education. International large-scale studies also assess how schools in participating countries enhance students' reasoning skills. Reasoning is one of the three cognitive domains on TIMSS (Trends for International Mathematics and Science Study) (the other two are knowledge and application), with certain items focusing on reasoning (Martin, 2016). On PISA (Programs for International Student Assessment), innovative domains (beyond reading mathematics and science) measure students' general cognitive skills, such as complex problem-solving (OECD, 2004), creative problem-solving (OECD, 2014) and collaborative problem-solving (OECD, 2017c). Science education is an especially important domain of school learning, which offers outstanding opportunities for improving student's specific and general cognitive skills; conversely, a certain level of reasoning skills is essential for understanding science learning materials. In the present study, we explore the latter aspect of learning sciences and the decisive role of reasoning skills. Taking into account the complexity of influencing factors, we also include further variables, such as motivation and parent involvement.

Children's academic success derives from an effort-making process involving a number of interactive factors including personal characteristics (demographic and personality-related), student motivation, and family and school care (Hellas et al., 2018). Cognitive abilities, such as inductive reasoning (IR) and scientific reasoning (SR), are inborn and enable students to be shaped in long-term educational contexts (Adey & Csapó, 2012; Adey & Shayer, 1994; Ariës et al., 2016; Boujaoude et al., 2007; Kyllonen et al., 2019; Molnár, 2011; Venville & Oliver, 2015). Cognitive development in children results from education and care from early childhood (Burger, 2010) and is influenced by stimulation both in school programmes and daily activities (Boroş & Sas, 2011; Resing et al., 2017). Previous studies have demonstrated the importance of IR in learning in most school disciplines (e.g. Adey, & Csapó, 2012; Hamers et al., 1998); it is closely tied to SR and problem-solving skills (e.g. Molnár et al., 2013; Schweizer et al., 2013). Cross-sectional data at different grade levels can offer a deeper understanding of development in reasoning, which plays a principal role in enhancing children's academic achievement and the reform process in particular contexts.

Additionally, previous studies have shown that students with greater motivation to reach their aims are more likely to succeed than those who are more talented but do not set their own goals and keep focusing on them (e.g. Duckworth et al., 2011; Eccles & Wigfield, 2002; OECD, 2017b). In learning science, motivation is not only the main factor in explaining science attitude, but also an essential predictor of science learning performance (Chan & Norlizah, 2018; Patrick et al., 2009) and academic achievement (Cavas, 2011; Dermitzaki et al., 2013). Children's activities in school and family environments have an influence on science motivation (SM) and achievement (Pintrich & Schunk, 2002). For instance, experiences in a school environment (Bathgate & Schunn, 2017; Hernesniemi et al., 2020) and parental involvement in schooling (Fan et al., 2012; Gonzalez-DeHass et al., 2005; Van Vo & Csapó, 2021a) have a positive effect on children's motivation in learning.

However, the research on links between cognitive abilities, motivation and parental involvement in predicting school achievement is rather scarce. Previous studies (e.g. Kriegbaum et al., 2018; Steinmayr & Spinath, 2009) have investigated the interaction between intelligence and motivation in predicting academic achievement. It seems that motivational factors have an impact on academic performance beyond intelligence (Steinmayr & Spinath, 2009). Studies on the family-school relationship have also confirmed the importance of cooperation between students' parents and school for children's academic success (Gonida & Urdan, 2007). Thus, this cross-sectional study aims to investigate the changing patterns of IR, SR and SM across grade levels. Gender differences in reasoning and SM are also examined in different grade cohorts. Furthermore, we explore the interactive factors of IR, SR and SM with the parental involvement variable in predicting students' STEM achievement.

5.2 Research questions and hypotheses

The foci of our study are to explore the changing patterns and relationship between IR, SR and SM through grade cohorts. We also investigate the extent to which IR, SR, SM and parental factors in predicting the STEM performance in secondary school students. Hence, our adapted test instruments were tailored to address following four research questions and hypotheses:

RQ2.1: What is the evidence for the reliability and validity of the adapted instruments?

H2.1: We expect that the psychometric properties are good for the IR test (Csapó et al., 2019; Kambeyo, 2018; Molnár & Csapó, 2011; Wu & Molnár, 2018) and acceptable for the SR test (Kambeyo, 2018; Korom et al., 2017) and for the SMTSL questionnaire (e.g. Cavas, 2011; Chan & Norlizah, 2018; Dermitzaki et al., 2013; Shaakumeni & Csapó, 2018; Tuan et al., 2005).

RQ2.2: How do students' IR, SR and SM performance differ in different school grades?

H2.2: Students in the older cohorts are expected to perform better than those in the younger ones on the IR test (e.g. Csapó, 1997; Díaz-Morales & Escribano, 2013; Molnár et al., 2013; Muniz et al., 2012) and on the SR test (Ding, 2018; Korom et al., 2017; Kwon & Lawson, 2000; Tairab, 2015), while students' motivation toward learning science is assumed to be a slight reduce through the lower grades to the upper grades (e.g. Dorfman & Fortus, 2019; Józsa et al., 2017).

RQ2.3: Is there any difference between male and female students in IR, SR and SM?

H2.3: It is expected that there is a non-significant difference between males and females in IR (e.g. Kambeyo, 2018; Molnár, 2011; Salihu et al., 2018) and SR (e.g. Mayer et al., 2014; Piraksa et al., 2014; Thuneberg et al., 2015), but females are hypothesised to be significantly more motivated towards science learning than males (Cavas, 2011; Chan & Norlizah, 2018; King & Ganotice, 2014).

RQ2.4: To what extent do IR, SR, SM and parent involvement in schooling variables interact in predicting STEM performance?

H2.4: It is supposed that IR, SR, SM and parental factors are main predictors of STEM performance (e.g. Coletta & Phillips, 2005; Lawson, 2000; Salihu et al., 2018).

5.3 Method

5.3.1 Participants

The study assessed 726 students (boys: 47.9%; girls: 52.1%) in six public schools in the southern province of An Giang (Vietnam). The average age of the participants was 14.0 years, with students in 19 classes. As presented in Table 5.1, the study sample included 4 grade cohorts in the 6th, 8th, 10th and 11th grades. We assessed the 11th graders instead of the 12th grade students, because it is difficult to approach the final year of secondary schools in the case. Each cohort consists of at least four classes from two different schools. The main data was collected between September and October of 2019. The students completed the test instruments in 45 min under the supervision of their teachers and our assistant teachers. The students took the tests and questionnaires in either paper-and-pencil or online versions. The effects of media administration modes were further discussed in Chapter 7.

Table 5.1. The participants of the study.

Grade	n	Boy/Girl ratio (%)	Mean age (years)	Age range (years)	No. of classes
6	139	43.9/56.1	11.3	10.8 - 12.7	4
8	260	51.5/48.5	13.3	12.8 - 13.9	6
10	166	49.4/50.6	15.3	14.2 - 15.9	5
11	161	44.1/55.9	16.4	15.8 - 17.4	4
All	726	47.9/52.1	14.0	10.8 - 17.4	19

5.3.2 Instruments

Background Questionnaire: The aim of the questionnaire, which adapted from the student questionnaire in PISA 2015 round (OECD, 2015), is to collect data on the students' background, parents' education level and parental involvement. The students were also asked to respond to items by scaling their perceptions of their parents' involvement in terms of parental support, engagement and interests in school activities (see Section 3.2.23).

Inductive reasoning test: To measure IR, we selected 20 items from the item bank, developed by the Research Group on the Development of Competencies of the University of Szeged (e.g., Csapó, 1997; Pásztor, 2016; Pásztor et al, 2017). The original items cover four subtests: figure series completion, figure analogies, number series, and number analogies (Korom et al., 2017). The selection of item criteria was based on the construction and psychometric characteristics of each item and empirical evidence from the previous studies, including an early study (See Chapter 4) in Vietnam (Van Vo & Csapó, 2020). Finally, the IR test consists of 20 items with five items for each subtest (see Section 3.2.1.1).

Scientific reasoning test: We used a 17-item test to measure SR, including main problem tasks: conservation, classification, proportional reasoning, correlational reasoning and IR tasks with science-related content materials. The content knowledge of the items related to basic concepts of science subjects in secondary schools. Most of the items were adapted and translated into Vietnamese from the original test by Korom et al. (2017) (see Section 3.2.1.2).

SMTSL questionnaire: The questionnaire is adapted from Tuan, Chin and Shieh (2005), that measures five motivational factors of SM (self-efficacy, science learning value, learning strategies, individual learning goals and learning environment stimulation). The validity of the questionnaire initially discussed in Chapter 4. In this study, we employed 18 items in 5-point Likert format (see Section 3.2.2.1).

5.3.3 Procedure and data analysis

The draft of the SR test were reviewed by two experts and three secondary school teachers before we conducted the field study. They focused on checking for any language issues and the relevance of discipline content in the test items. Three secondary school teachers also verified the relevance of the content knowledge in the test with the current program in the secondary education level in Vietnam. A pilot study was carried out with the final version among seven students in two public schools (one secondary school and one high school). We observed these students while they did the test. The items were then discussed and slightly adjusted related to language and visualization issues before conducting the main study.

The testing process was administered in either PP or online modes, depending on particular conditions of the participating schools. Students completed the instrument in 45 min. The PP mode was administered in the classrooms, and each student received a test booklet and an answer sheet. Teachers guided the students step by step following our instructional procedure. For the online testing, students access the eDia platform via a link and an individual code (Csapó & Molnár, 2019; Molnár & Csapó, 2019). Two teachers provided technology support and observed the students during the testing process. The online instruments were operated through the University of Szeged servers.

The results were scaled in the Rasch model with ACER ConQuest software in dichotomous items for the IR test (Adams & August, 2010) and in polytomous items with PCM analysis for the SMTSL questionnaire (Adams & Wu, 2010). Our raw scores were then converted to MLE scale as output parameters in the Rasch model measurement. The common packages in R programme version 3.5.3 (R Core Team, 2019) were used in this study: psych (Revelle, 2019), lavaan package (Rosseel, 2012) and yarrr (Phillips, 2016).

5.4 Results

5.4.1 Reliability and validity of the instruments

As emphasized in the purposes of the present study, we focused on interpreting the results of the test instruments in the unidimensional model in general. Reliability estimates based on the internal consistency indicator of McDonald's omega (ω) since it seems less biased than Cronbach's alpha in this case (Dunn et al., 2014). Omega values for the IR test, SR test and SMTSL questionnaire were .82 (Cronbach's alpha = .80), .64 (Cronbach's alpha = .61) and .90

(Cronbach’s alpha = .88), respectively, implying that they are acceptable in terms of internal consistency reliability.

The unidimensional Rasch model analysis also confirmed that all the items on the tests fitted the data quite well. According to Griffin (2010), an item is considered as an infit one when its infit value falls within a range of 0.7 and 1.3. In this study, the infit for individual items (weighted mean squares, MNSQ) ranged from 0.87 to 1.12 (M = 0.99, SD = 0.07) on the IR test (See Appendix E) and from 0.91 and 1.10 (M = 1.00, SD = 0.04) on the SR test (See Appendix F). There were 19 out of 20 items on the IR test, and 15 (out of 18 items on the SR test have got the discrimination value higher than 0.3, denoting that the quality of test items is acceptable (Ebel & Frisbie, 1991).

Likewise, we scaled the students’ performance on the SMTSL questionnaire in polytomous items with PCM in a unidimensional Rasch measurement, and its output was a science motivation score in general (Dermitzaki et al., 2013) (see Chapter 4, Section 4.4.6). The results showed that items fit the data well, with the infit values ranging from 0.82 to 1.43, but one item had the infit value higher than 1.3 (See Appendix G).

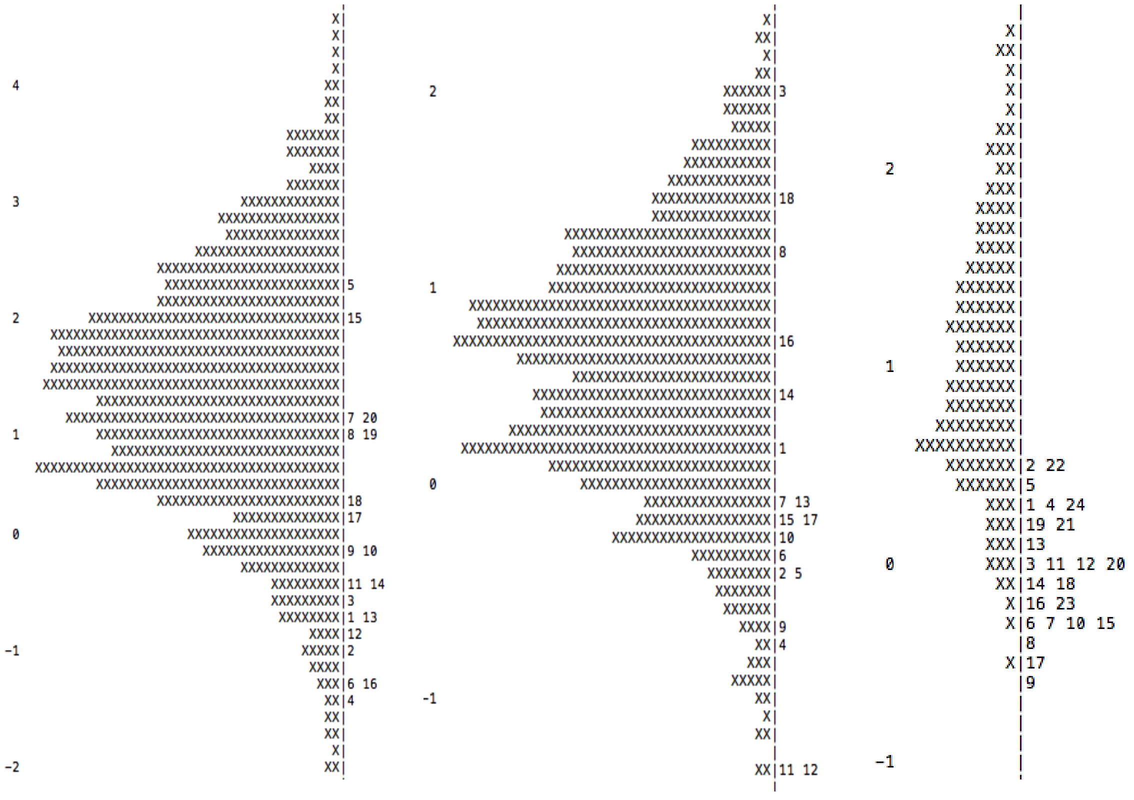
Furthermore, we used a DIF analysis to examine statistically invariant characteristics at item level. The R difR package (Magis et al., 2010) was employed for DIF analysis in dichotomously scored items with the Angoff's Delta method (see Angoff, 1982). The results showed that no DIF item was found on either test as regards gender or administration modes. For the SMTSL questionnaire, we conducted a DIF analysis to test measurement invariance with respect to gender and delivery modalities with the R lordif package (Choi et al., 2011). McFadden’s R^2 is set at 0.02 as a criterion for rejecting the null hypothesis of no DIF. All pseudo R^2 values were below 0.02 and p-values of the goodness-of-fit statistics above 0.05, indicating that no DIF items were detected following these criteria on all questionnaire items regarding gender and administration modes.

5.4.2 Descriptive statistics

On the IR test, the item difficulties ranged from −1.35 to 1.94 with the average item difficulty at 0 (as defaulted in the program). On average, the person proficiency of 1.36 logits (SD = 1.31) compared to the item difficulty of 0 logits, implying that the students performed higher than the item difficulty mean. Item difficulty values on the SR test ranged from −1.35 to 1.94, while the average person proficiency was 0.62 logits (SD = 0.89), suggesting that the students’ ability was higher than the item difficulty in average. Overall, participants correctly completed

71.75 % of the items (14.35 out of 20 items) on the IR test and 62.17 % of the items (11.19 out of 18 items) on the SR test.

The Wright maps were drawn to visualize estimation between item difficulty and the chances of each person's success by comparing the relative positions of the items and persons. The items are presented from easiest (bottom) to most difficult (top) on the right side of the map, and the left side of the map shows the distribution of the participants. As illustrated in Figure 5.1, a majority of the students showed high proficiency in reasoning and high motivation toward learning science. In fact, most of the students felt that the IR test was easier than the SR test, so more students achieved higher-performance scores on the IR test than they did on the SR test. The most difficult items were item 5 on the IR test and item 3 on the SR test, while item 4 on the IR test and item 11 on the SR test seemed the easiest ones. The spectrum of the students' ability on the IR test was wider than that of students' ability on the SR test. However, item difficulty distributed the most area of the scale and covered all of test-takers' ability, indicating that the tests appeared fairly to measure reasoning proficiencies among the participants.



(a) Inductive reasoning test (b) Scientific reasoning test (c) SMTSL questionnaire

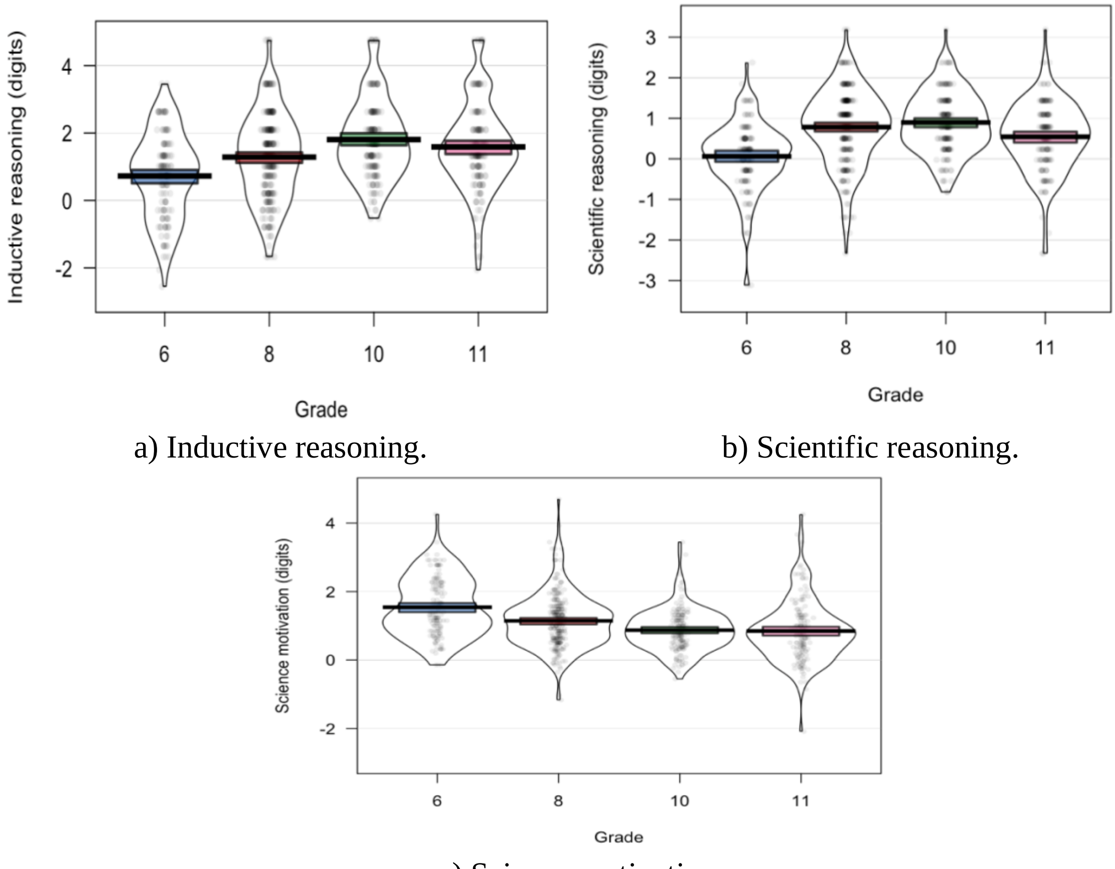
Figure 5.1. Wright map of latent distributions and response model parameter estimates.

Note. Each 'X' represents (a) 0.1 cases, (b) 0.9 cases and (c) 4.1 cases.

5.4.3 Performance of students in different cohorts on the tests and questionnaire

The pirate plots were used to depict the performance of participants on the IR, SR tests and SMTSL questionnaire in different cohorts. Figure 5.2 showed the pirate plots illustrating the students' attainment on the (a) IR test, (b) SR test and (c) SMTSL questionnaire.

For the IR test, the students' mean scores increased remarkably from the 6th (M = 0.73, SD = 1.25), to the 8th (M = 1.29, SD = 1.29) and on to the 10th grades (M = 1.81, SD = 1.17), before decreasing slightly in the 11th grade (M = 1.59, SD = 1.32) (see Figure 5.2a). In the 6th grade, although the mean score was lower than that of other grades, some of the students had high proficiency on the IR test. Both the highest- and lowest-performing participants seemed to be in the 8th grade, and the distribution of scores was balanced in two directions of distribution shape. On average, the 10th graders yielded a highest score, and most of the students earned more than 0 with the lowest standard deviation. The bean density in the pirate plot for the 11th grade was slightly different from that of other grade groups. Several students achieved scores of around 3.0 (logits), but some students in the 11th grade were on the lowest-score list.



a) Inductive reasoning.

b) Scientific reasoning.

c) Science motivation.

Figure 5.2. Students' performance on the (a) IR test, (b) SR test, and (c) SMTSL questionnaire across grade levels.

In the same trend in IR, the students in the older groups performed better than their younger counterparts on the SR test, except in the 11th grade (Figure 5.2b). The 6th graders had an average score of 0.06 (SD = 0.84), but some students received the lowest points. The students in the 8th grade gained a mean score of 0.78 (SD = 0.92), a dramatic rise compared to the 6th graders. Most students who earned the highest scores were in the 8th and 10th grades. The 10th graders achieved the highest mean points (M = 0.90, SD = 0.75), but an increasing trend seemed to reverse in the 11th grade (M = 0.55, SD = 0.84).

In contrast, the students' motivation toward learning science tended to drop slightly across the grade levels (Figure 5.2c). The 6th graders achieved the highest score (M = 1.54, SD = 0.86), followed by the 8th graders (M = 1.15, SD = 0.79). However, the students' motivation followed the same pattern between at the 10th (M = 0.87, SD = 0.61) and the 11th grades (M = 0.84, SD = 0.88), but no 10th graders fell into the lowest-score group, while some of the 11th graders scored so high and or achieved so low on the SMTSL questionnaire.

In addition, we employed the ANOVA analysis to explore the effect of grade levels on the students' reasoning proficiency and motivation. The results showed that there was a significant difference between the grade cohorts on the IR test [F(3, 722) = 20.53, p < .001], SR test [F(3, 722) = 29.52, p < .001] and SMTSL questionnaire [F(3, 722) = 25.1, p< .001]. Tukey's HSD was implemented as *post-hoc* analysis to test significant differences in pairs of grades. As summarised in Table 5.2, the older groups earned higher scores than the younger ones on two reasoning tests, except for the 11th grade. The 8th graders showed a significant improvement compared to the 6th graders on both the IR and SR tests. The 10th graders also achieved higher scores than those in the 8th grade, but no significant difference was found between these two grades. Surprisingly, the 11th graders performed significantly lower than the 10th graders on the SR test. Generally, the students in all the upper grades were significantly less motivated than their juniors, excepted between the 11th and 10th, where the 10th and 11th graders showed the same level of SM.

Table 5.2. Tukey's HSD in multiple comparisons between grade cohorts.

Grade	Inductive reasoning		Scientific reasoning		Science motivation	
	ΔM	*Cohens' d*	ΔM	*Cohens' d*	ΔM	*Cohens' d*
8 & 6	0.56	0.44***	0.721	0.81***	−0.40	0.49***
10 & 8	0.52	0.42 ***	0.114	0.13	−0.27	0.37**
11 & 10	−0.22	0.17	−0.352	0.44**	−0.03	0.03

Note. ΔM: Mean difference.
**p< .01, ***p< .001.

5.4.4 Gender difference in IR, SR and SM

Figure 5.3 demonstrates the differences in the students' performances on the reasoning tests and SMTSL questionnaire throughout the grade cohorts. In general, the changing curves for each gender were nearly the same pattern between the two reasoning tests. However, there was a slight difference in cognitive development between males and females. The girls appeared to develop reasoning abilities earlier than the boys, but after the middle school years, the difference remained unchanged or even dropped slightly in the girl group, while boys showed an ongoing improvement until the end of the first year in high school. Unlike reasoning skills, the general trend of student motivation reduced gradually across grade levels. Girls in the 6th grade seemed more motivated than boys, but boys had a higher mean score than the girls in the 11th -grade group.

Furthermore, we manipulated the *t*-test to compare IR, SR and SM between males and females. No significant difference was found between males and females on either the reasoning tests or the SMTSL questionnaire for the entire sample (Table 5.3). It also showed the same trend on the reasoning test in each grade cohort, except in the 10th grade, where males performed significantly better than females. For the SM test, no significant difference was indicated in the 8th grade or 10th grade, but motivation scores differed significantly regarding gender in the 6th grade (in favour of girls) and at the 11th grade (in favour of boys).

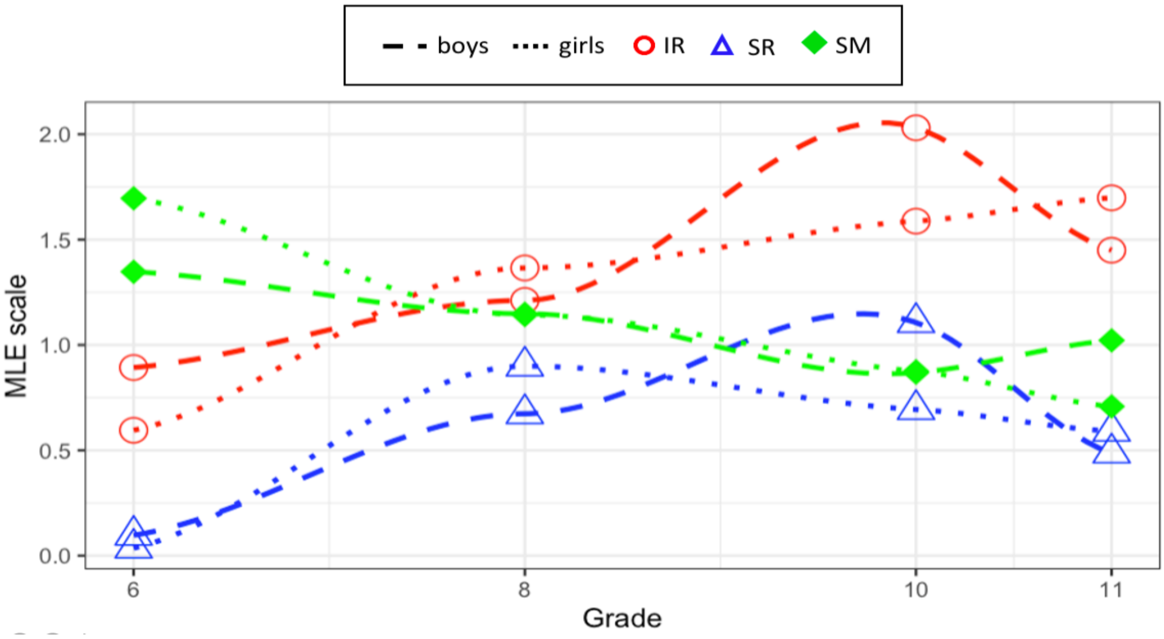


Figure 5.3. Comparison of performance of students in different grade cohorts.

Note. IR: inductive reasoning; SR: scientific reasoning; SM: science motivation.

Table 5.3. The *t*-test for comparing performance between boys and girls in IR, SR and SM.

Grade	Inductive reasoning			Scientific reasoning			Science motivation		
	ΔM	p	Cohens' d	ΔM	p	Cohens' d	ΔM	p	Cohens' d
6	.30	.18	0.24	.06	.68	0.07	−.42	.02	0.41
8	−.15	.34	0.11	−.23	.05	0.24	.00	.95	0.01
10	.44	.02	0.38	.41	<.01	0.57	.00	.95	0.01
11	−.25	.23	0.19	−.10	.44	0.12	.31	.04	0.35
All	.06	.53	0.05	.03	.61	0.04	−.00	.97	0.00

Note. ΔM: mean difference between boys and girls.

5.4.5 Predicting STEM achievement

We applied a path analysis in the R lavaan package (Rosseel, 2012) to investigate the relevant predictors explaining the students' STEM achievement. In this study, we proposed the exploratory variables, including inductive reasoning (IR), scientific reasoning (SR), science motivation, mother's education level (ME), father's educational level (FE) and parental involvement in schooling (PI) in predicting the STEM achievement (STEM). We employed the mean score of four school subject tests in mathematics, physics, chemistry and biology in the last term as an index of the STEM performance. It may range from 0.0 to a 10.0 based on a 10-point scale that has officially been introduced in secondary schools in Vietnam. In the study, the results of these tests were collected using the self-report form in the background questionnaire. As suggested in the literature, we proposed a hypothesized model as depicted in Figure 5.4.

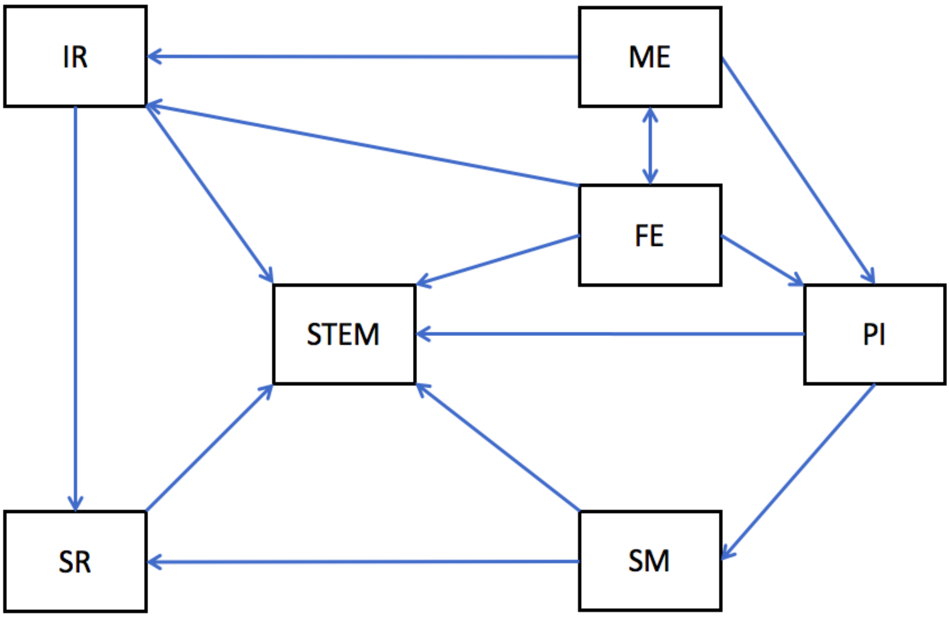


Figure 5.4. A hypothesized model for predicting the STEM achievement.

Note. IR: inductive reasoning; SR: scientific reasoning; SM: science motivation; PI: parental involvement in schooling; ME: mother's education level; FE: father's education level.

After conducting several investigations, the final model in Figure 5.5 with six predictors, presenting significant relations between factors was demonstrably the best one. The results confirmed that the model was a good fit to the empirical data, with: $\chi^2(12) = 24.12$, p = .02, CFI = .979, TLI = .964, SRMR = .047, RMSEA = .050. The model shows that among six the proposed predictors, only four ones (inductive reasoning, scientific reasoning, science motivation, and father's education level) have a direct effect and can explain around 25.5% of the STEM achievement variance. Unexpectedly, ME does not contribute directly to explaining STEM achievement with statistical significance, but this variable has a strong link to FE and a direct effect on PI. In contrast, FE is the best factor in directly predicting STEM achievement, followed by SM and SR. IR was shown to be the least direct predictor of STEM achievement, but it also had an indirect effect on STEM achievement through the SR variable. This mediated relation was examined with the Sobel test, with the determination that the mediating effect was significant (Z = 3.71, p < .001). In spite of not being a direct predictor, PI has an indirect effect on STEM achievement via the mediator of the SM variable, based on the Sobel test (Z = 3.60, p < .001).

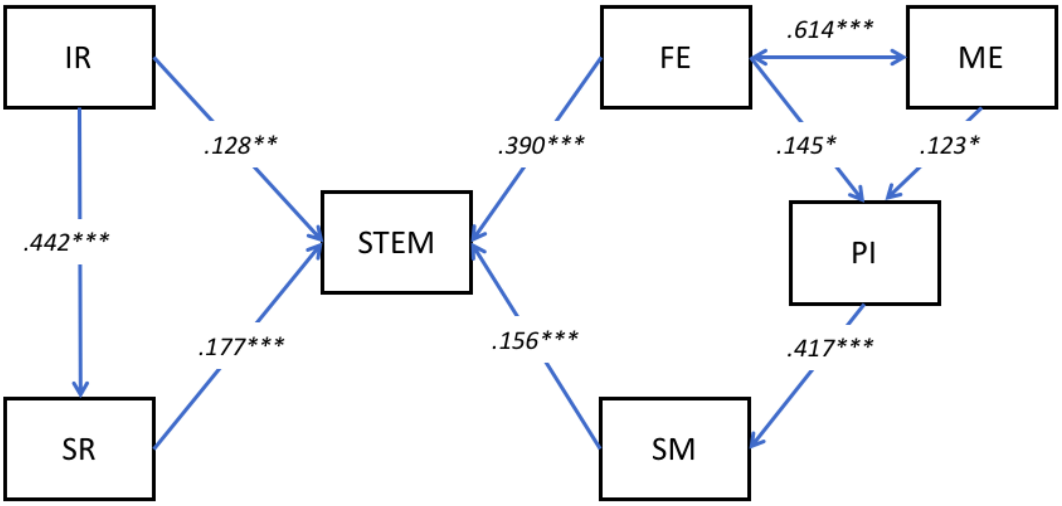


Figure 5.5. The proposed model for predicting the STEM achievement.

Note. IR: inductive reasoning; SR: scientific reasoning; SM: science motivation; PI: parental involvement in schooling; ME: mother's education level; FE: father's education level.

*p < .05, **p<.01, ***p<.001.

5.5 Discussion and conclusion

The validity and reliability of the adapted tests (hypothesis H2.1) have been confirmed, with these instruments shown to be potential tools for assessing reasoning students' proficiency and SM in the Vietnamese context. However, a few items (e.g., item 3 and item 18) in the SR test should be reconsidered for future tests. The findings are in line with previous studies (Dermitzaki et al., 2013; Kambeyo, 2018; Korom et al., 2017; Van Vo & Csapó, 2020). In general, the students' proficiencies were estimated higher than the average item difficulty on the two reasoning tests. In next-generation tests, more difficult items should be involved to measure high-performance students.

The students' performance on both the IR and SR tests increased gradually across grade levels. Specifically, the students' reasoning improved remarkably through the lower secondary education level, but the growth rate appeared to reduce in the upper grade cohorts. The hypothesis H2.2 is partially verified. The results are somewhat consistent with previous studies on IR (Csapó, 1997; Díaz-Morales & Escribano, 2013; Molnár et al., 2013; Van Vo & Csapó, 2020) and on SR (Ding, 2018; Kwon & Lawson, 2000; Tairab, 2015). Nonetheless, an inconsistent finding in the current study is that the rate of the decrease in the period of the 10th - 11th grades was higher, while previous studies recorded a slight rise between those grades. The findings partly reflect that the lower secondary curricula seem to be more successful in enhancing the students' optimum development in thinking skills rather than the upper secondary curricula. Interestingly, although the children achieved higher scores on the IR test, the developmental trend appears to be similar on both the two tests. This provides evidence of a close link between the two forms of reasoning skills, in which teaching reasoning through content knowledge in individual subjects can contribute to developing the general thinking skills in children. In contrast with reasoning abilities, the students' motivation toward learning science dropped gradually through the grade levels. The findings are mostly consistent with previous research (e.g. Bouffard et al., 2001; Dorfman & Fortus, 2019; Józsa et al., 2017), but the changing curve seemed to slightly reverse in high school students.

As regards gender, no significant difference was found on either the reasoning tests or the SMTSL questionnaire, but the mean scores for males was slightly higher than those of females on the reasoning tests, and the mean score for females on the SMTSL was rather higher than that of males. These findings are consistent with previous studies in the same context in Vietnam (Van Vo & Csapó, 2020, 2021) (Hypothesis H2.3 is corroborated). A significant difference was only indicated in the 10th grade. A possible reason might be related to the shift from the lower secondary education to upper secondary education when learning in a new learning environment and in a new learning program might more or less influence the students' reasoning capacities. However, the reasoning abilities in girls appeared to develop earlier than those of boys in early secondary education, with the growth rate slowing among girls when they entered high schools. The effects of gender on cognitive development in IR and SR offer some helpful clues to optimize enhancing thinking skills in school practice. Teachers have recognised the trend and advocate finding more appropriate facilities to maintain student improvement, especially among female students after the middle school years.

The path analysis showed that cognitive abilities, motivation and parental factors are important predictors of STEM success. The results confirm the hypothesis H2.4. The findings are fairly consistent with existing studies (e.g. Fan & Williams, 2010; Steinmayr & Spinath, 2009) and a meta-analysis review by Kriegbaum et al. (2018) as regards the core roles of intelligence and motivation in predicting academic attainment. Parental factors also contribute meaningfully to the students' academic success. Specifically, the father's education level was found to be the main predictor of STEM achievement, while the mother's education level indirectly influenced children's motivation through parental involvement in schooling. This is because parents' education level is an indication of the educational resources children have at home (OECD, 2017a). Furthermore, parental involvement in schoolwork is a significant factor in predicting children's success in learning STEM subjects. These results are consistent with

Ganzach (2000), and they may be explained by the culture in Vietnam, where parents tend to focus intensely on their children's performance in school (Hoang et al., 2014; Phan, 2004).

In line with Kambeyo (2018) and Korom et al. (2017), there is a strong, positive correlation between IR and SR test scores. The current study investigated the relationship between IR general thinking skills and SR in the domain-specific content of science in Vietnam. Thinking can be taught directly in specific courses or embedded in the regular school curricula within the framework of school disciplines (Csapó, 1999). In Vietnam, it is encouraged to integrate thinking skills into the teaching of individual disciplines. The study showed that the development of thinking skills can be monitored with the infusion approach. The difference in children's reasoning between lower and upper secondary schools may come from the biological development of children or from different curricula between two education levels. This issue should be considered in future research.

CHAPTER 6. ASSESSING SCIENTIFIC REASONING IN CONTROL OF VARIABLES STRATEGY AND STUDENTS' MOTIVATION IN LEARNING PHYSICS IN SECONDARY SCHOOL STUDENTS

6.1 Introduction

In modern society, there is increasing pressure to manage more information in a shorter time. Mastery of scientific skills is a core goal of science curricula around the world. However, students should be trained in general thinking skills of information processes rather than instructed in extensive content knowledge in individual subjects. TIMSS and PISA have considered SR the main component of science literacy assessment (Osborne, 2013). CVS is a leading SR skill that helps learners conduct, predict and evaluate experiments during the learning process (Chen & Klahr, 1999). The ability to construct arguments and understand principles of unconfounded evidence is essential in science education. The developmental psychology of CVS relates directly to the development of SR ability (Kuhn, 2007). CVS is foundational for science and scientific literacy, which could be promoted in school contexts (Schwichow, Croker, et al., 2016). Assessment of CVS can clarify children's cognitive development in applying reasoning skills and interpreting experiments across various age groups.

In Vietnam, critical thinking and reasoning are implicitly embedded in core curricula; experimental and inquiry activities have been compulsory for science teachers. Students learn science in separate subjects beginning in the first year (6th grade) of secondary education. Experimental activities account for a major proportion of the science subjects' programs. Particularly, the national education program requires around 10% of the physics curriculum in high schools to involve experimental lab activities (MOET, 2009).

However, students' CVS proficiency rarely seems to be assessed on tests at the school and state levels or even on the national examination in Vietnam. Assessment of CVS is one means to evaluate the effects of current curricula on developing students' SR abilities and encourage schools to develop the students' SR proficiency. More understanding of children's cognitive development plays an important role in improving school practice and reforming the next-generation curricula. This study attempts to develop and validate the test as an instrument to measure CVS in physics for secondary school students. Our instruments mainly focus on assessing the ability of students to identify, interpret and understand whether a variable affects the behavior of an experimental set. In addition, this study also outlines basic rationales of assessing CVS, exploring latent factors predicting item difficulty as a practical reference for teachers who consider designing assessment-based learning activities in classrooms and hands-on tasks in school labs. We also investigated the effects of age-group (grade level), gender and content knowledge on the CVS proficiency in secondary school children. Furthermore, the development of CVS capacity was observed in its relationship with motivation across grade levels. Proposed findings were expected to provide insight into developmental patterns of CVS in school-age children, which may offer useful information for school practice.

6.2 Research questions and hypotheses

The main purposes of this study are to develop a CVS test that involves three sub-skills (ID, IN and UN) of CVS with the contents related to basic physics in secondary education level, to examine its psychometric characteristics, and to investigate the developmental patterns of students' scientific reasoning in CVS ability and motivation in physics. We also explore the latent factors that contribute to explaining the individual CVS capacity in Vietnam context. Thus, the present study addresses the following research questions and hypotheses:

RQ3.1: What is the structure of the test developed to assess secondary school students in the Vietnamese context?

H3.1: We assumed that the proposed test involves three subskills (ID, IN and UN) of CVS (Chen, & Klahr, 1999) with the contents related to basic physics in secondary education level.

RQ3.2: What is the evidence for validity and reliability of the CVSP test?

H3.2: We expect that it is an acceptable level in terms of psychometric properties in the CVSP test (Schwichow, Christoph, et al., 2016).

RQ3.3: Which latent factors (e.g., subskills, physics-related contents and number of independent) impact item difficulty in the CVSP test?

H3.3: It is assumed that sub-skills of CVS influence on item difficulty (Schwichow, Christoph, et al., 2016), but item difficulty is not impacted by physics-related contents difficulty (Schwichow, Christoph, et al., 2016) or number of independent (Staver, 1986).

RQ3.4: What is the evidence for validity and reliability of the adapted student motivation in learning physics questionnaire?

H3.4: It is expected that the adapted questionnaire is a reliable and valid measurement tool for assessing student motivation toward learning physics in Vietnam context (Hooper et al., 2013).

RQ3.5: To what extent do students' capacity in CVS develop across grade levels?

H3.5: We expected that students in the older cohort perform significantly better than students in the younger groups (Han, 2013; Schwichow et al., 2020; Zimmerman, 2007).

RQ3.6: Do students' motivation toward physics learning differ in different grade cohorts?

H3.6: It is assumed that there is no significant difference in student motivation among grade levels.

RQ3.7: Is there a different performance between males and females on the test and questionnaire?

H3.7: We suppose that no significant difference is found between males and females in the CVSP test (Mayer et al., 2014; Piraksa et al., 2014) and the physics motivation questionnaire.

RQ3.8: To what extent are interaction of physics motivation and physics content variables in predicting students' CVS proficiency in physics?

H3.8: We expect that motivation and the physics content test are closely associated with CVS ability (Schwichow et al., 2020; Zhang & Bae, 2020) and contribute to explaining individual CVS capacity.

6.3 Procedure and method

6.3.1 Development of the control of variable in physics test

Chen and Klahr (1999) classified CVS into four subskills: planning controlled experiments (PL), identifying controlled experiments (hereafter referred as "Identifying" or "ID"), interpreting controlled experiments (hereafter referred as "Interpreting" or "IN") and understanding the indeterminacy of confounded experiments (hereafter referred as "Understanding" or "UN"). PL requires participants to design an experimental system from given materials in which variables are constrained. This task aims to produce a proposal plan in which students need to decide which variables to manage and which to observe. In ID tasks, students must distinguish between testability and causal influences of variables in an experimental system to decide which suitable variables should be included to reach a particular conclusion. A problem often begins with a stem, such as a short story that someone needs to prove or test a hypothesis of a causal association. Participants will choose one among some alternative options that are presented in the given experimental sets. Only a controlled experiment is correct, while other distractors represent confounded experiments with more than one variable changed. When interpreting controlled experiments, the IN task reveals students' abilities in inferring the outcome of a controlled experiment. A stem of this problem task shows the possible outcomes of experimental systems. Students who know how to draw suitable inferences from a controlled experiment to reach valid outcomes are more likely to get a correct answer. Participants must prove their profound understanding of the indeterminacy of confounded experiments. The stem structure in UN task is similar to IN one, but distractors are confounded, or valid outcomes drawn from an experimental system. To arrive at a correct answer, students have to understand the components of the given experiment and decide whether they can draw valid conclusions from those outcomes.

We referred to the process guided by rigorous psychometric standards (Kane, 2016; Messick, 1995) to develop the test measuring three subskills of CVS in physics for secondary school students. Particularly, the CVSP test was composed of three subskills of CVS: Identifying, Interpreting and Understanding. The content of test items involves the knowledge related to physics, including basic concepts emphasized in the Vietnamese secondary educational curricula, such as mechanics, heat, thermodynamics, electricity and electromagnetism. These kinds of problem tasks cover both cognitive processes in terms of CVS and domain content in physics. Table 6.1 presents the contributions of the test items. Nine items were adapted from the test instrument developed by Schwichow, Christoph, et al. (2016). Five items were modified from the American Association for the Advancement of Science (AAAS) database and TIMSS (1997), while the authors developed ten new items. Table 6.2 provides the referenced sources of the CSVP test.

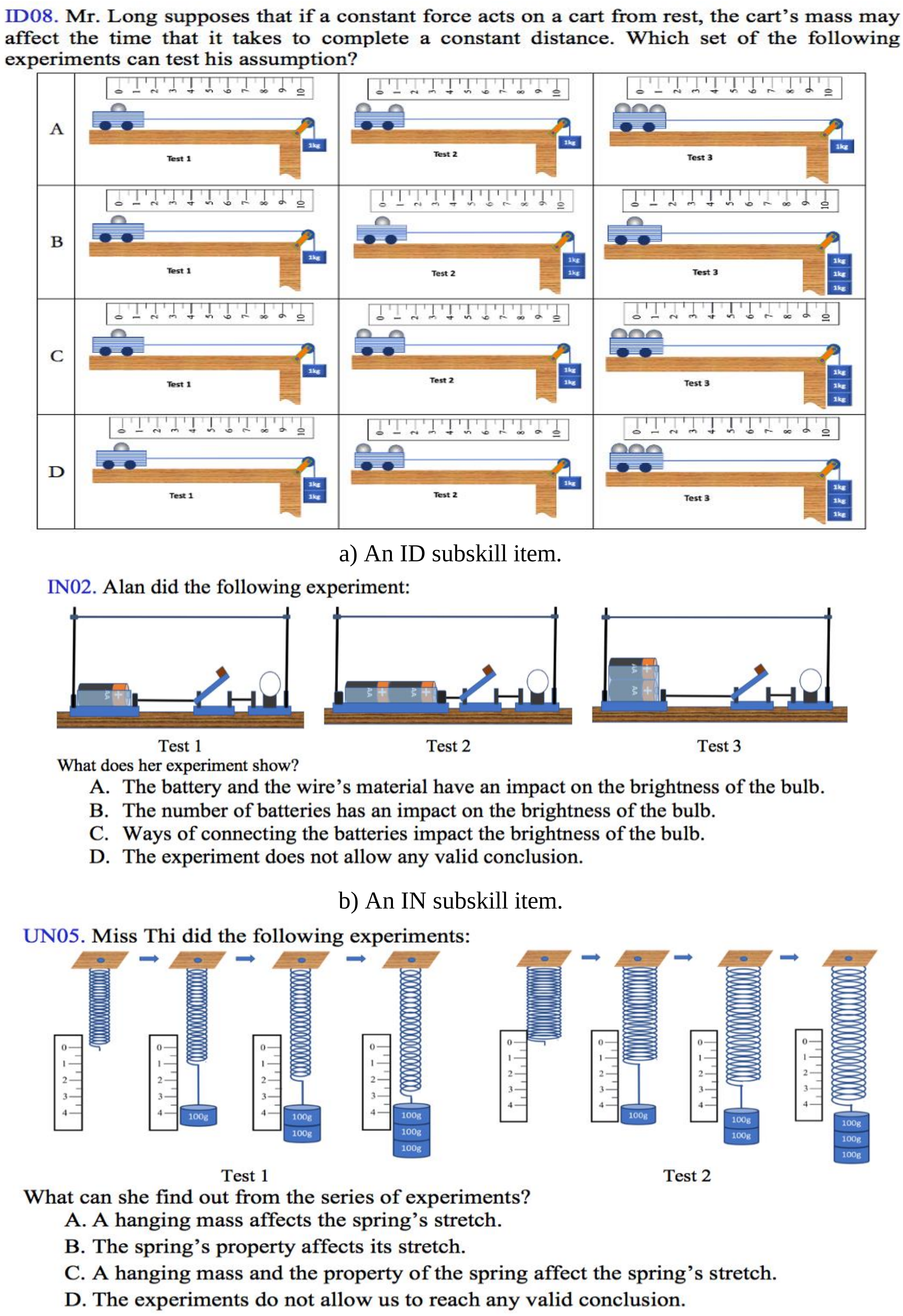
Table 6.1. Two-dimensional table of specification of the CVSP test.

Content	Identifying	Interpreting	Understanding	Total
Mechanics	3	2	3	8
Heat and Thermodynamics	2	2	3	7
Electricity and Electromagnetism	3	4	2	9
Total	8	8	8	24

We developed the test items in a multiple-choice format, using a stem with three distractors and one correct answer. Multiple choice has been the most popular method to measure scientific reasoning (Opitz et al., 2017), yielding smaller effect sizes than other instrument formats (Schwichow, Croker, et al., 2016). It can increase precision in assessing more respondents and facilitate better data collection and scoring than other alternative assessment formats. Graphical representations minimized the influence of students' varying reading ability levels. The test is constructed in three subskills of interpreting, identifying, and understanding. The test specification involves content coverage of basic physics curricula in Vietnam and the cognitive process dimension of three subskills of CVS. The first subskill, planning, is not involved in the current multiple-choice item test because of the limitations of the paper-based version. A correct and incorrect answer were initially scored as 1 and 0 points, respectively. Samples of the test items are presented in Figure 6.1 (see more Appendix D).

Table 6.2. Summary of the characteristics and referenced sources in the CVSP test.

Item	Content	Source referenced	Note
Identifying controlled experiments			
ID01	Heat and Thermodynamics	Schwichow, Christoph, et al. (2016)	
ID02	Mechanics	TIMSS (1997)	New image
ID03	Heat and Thermodynamics	Authors	
ID04	Heat and Thermodynamics	Schwichow, Christoph, et al. (2016)	
ID05	Electricity and Electromagnetism	Authors	
ID06	Electricity and Electromagnetism	Schwichow, Christoph, et al. (2016)	
ID07	Mechanics	AAAS	
ID08	Mechanics	Authors	
Interpreting controlled experiments			
IN01	Mechanics	AAAS	New image
IN02	Electricity and Electromagnetism	Authors	
IN03	Electricity and Electromagnetism	Schwichow, Christoph, et al. (2016)	New image
IN04	Mechanics	Authors	
IN05	Electricity and Electromagnetism	Schwichow, Christoph, et al. (2016)	New image
IN06	Heat and Thermodynamics	AAAS	
IN07	Mechanics	Authors	
IN08	Electricity and Electromagnetism	Authors	
Understanding the indeterminacy of confounded experiments			
UN01	Electricity and Electromagnetism	Schwichow, Christoph, et al. (2016)	
UN02	Heat and Thermodynamics	Schwichow, Christoph, et al. (2016)	New image
UN03	Heat and Thermodynamics	Schwichow, Christoph, et al. (2016)	
UN04	Mechanics	AAAS	
UN05	Mechanics	Authors	
UN06	Electricity and Electromagnetism	Schwichow, Christoph, et al. (2016)	
UN07	Mechanics	Authors	
UN08	Heat and Thermodynamics	Authors	



ID08. Mr. Long supposes that if a constant force acts on a cart from rest, the cart's mass may affect the time that it takes to complete a constant distance. Which set of the following experiments can test his assumption?

a) An ID subskill item.

IN02. Alan did the following experiment:

What does her experiment show?

A. The battery and the wire's material have an impact on the brightness of the bulb.
B. The number of batteries has an impact on the brightness of the bulb.
C. Ways of connecting the batteries impact the brightness of the bulb.
D. The experiment does not allow any valid conclusion.

b) An IN subskill item.

UN05. Miss Thi did the following experiments:

What can she find out from the series of experiments?

A. A hanging mass affects the spring's stretch.
B. The spring's property affects its stretch.
C. A hanging mass and the property of the spring affect the spring's stretch.
D. The experiments do not allow us to reach any valid conclusion.

c) An UN subskill item.

Figure 6.1. Examples of items in the CVSP test.

6.3.2 Investigations of the CVSP test

6.3.2.1 Initial content validity testing

The first draft of our proposed test was sent to two experts in physics education and three physics high school teachers to review. After modifying the language-translation issues and comparing the relevant content to the current curricula, the second version of the instrument was implemented as a pilot study with ten voluntary participants in a high school. We introduced the aims of our pilot work before students completed the test under our supervision. Following the trial test, students were asked about language issues and the knowledge contained in the test. The final version was adjusted to deal with these minor matters.

6.3.2.2 Participants

The main study was conducted with 807 students from the 8th to 12th grades in Vietnam (Table 6.3). The participants had a mean age of 15.5 years with a male-female ratio of 40.6:59.4. We randomly selected 21 classes in 11 schools. Students voluntarily joined the study after their teachers introduced the aims of our project. The data was collected in 2020 and 2021. Students took the test either in paper-and-pencil or online administration modes depending on the particular circumstances in each participating school. Measurement invariance with respect to delivery modalities was discussed in the Chapter 7.

Table 6.3. Study 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

The background questionnaire was embedded in the first section of the test in both versions. Students were also asked to report their GPA and the marks they earned on the final physics tests in the previous semester. In this study, we adapted the student questionnaire in TIMSS 2015 (Physics in school) (Hooper et al., 2013; TIMSS, 2015) to assess students' motivation and attitude in learning physics, with three main scales: student like learning physics, students' views on engaging in physics lessons, and students' confidence in learning physics (see Section 3.2.2.2). The online test was designed in the eDia via the servers of the

University of Szeged. In principle, students spent around 45 min completing the test and questionnaire.

6.3.2.3 Statistical analysis

Based on the psychometric guidelines of Kane (2016) and Messick (1995), and the Standards for Educational and Psychological Testing (American Education Research Association, American Psychological Association, 1999), the CVSP test development mainly involved investigating the reliability evidence, dimensionality of the test models, item-model fit in the Rasch model, and measurement invariance at the item level.

We employed the Rasch model in ACER ConQuest software with dichotomous items (Adams & August 2010) to convert the raw score data to a linear scale. Such outputs of the program provide the main parameters of model fit statistics and reliability, item-model fit, item ability distribution fit, and some main classical statistics indices as well. Regarding dimensional model fit, the efficacy of the possible models in Rasch model analysis is often based on three indices: the deviance, AIC and BIC. By comparing the final deviances, AIC and BIC, a model with the lower coefficients is a better fit. In each model, the cut-off standard for the acceptable value of the infit index is in the range of 0.77 - 1.30 (Griffin, 2010). An item is fitted with the Rasch measurement model if it meets this criterion. A good item fit denoted that the probability of students responding correctly on an item increases with their proficiency. A Wright map was also considered to explore the relationship between participants' responses and item difficulty (see Section 3.3.2).

In addition, ANOVA analysis and Tukey's HSD test were employed to examine the effects of subskills, knowledge content, and the number of independent variables on the item difficulty. We also investigated the association between CVS and content knowledge. In this case, we utilized the R psych package (Revelle, 2019) and ggplot2 package (Wickham, 2016) to tailor the outcomes of ANOVA and visualize the pattern of item difficulty in the test.

6.4 Results

6.4.1 Validity and reliability of the CVSP test

6.4.1.1 Unidimensional model and three-dimensional model

To explore the difference in the internal structure of the test, we fitted unidimensional and three-dimensional models to the dataset. A formal statistical test of these models' relative fit can be undertaken by comparing the final deviance of the two models. In the current study, the unidimensional model got the final deviance value of 22,048.707 with 25 parameters estimate

(AIC = 22,098.71, BIC = 22,121.38), and the three-dimensional model had 21,989.499 with 30 parameters estimated (AIC = 22,049.5, BIC = 22,2076.71). The deviances in the unidimensional model were greater than those in the three-dimensional model. Consequently, we can conclude that the three-dimensional model had a significantly better fit than the unidimensional one.

6.4.1.2 Item quality and participants' response

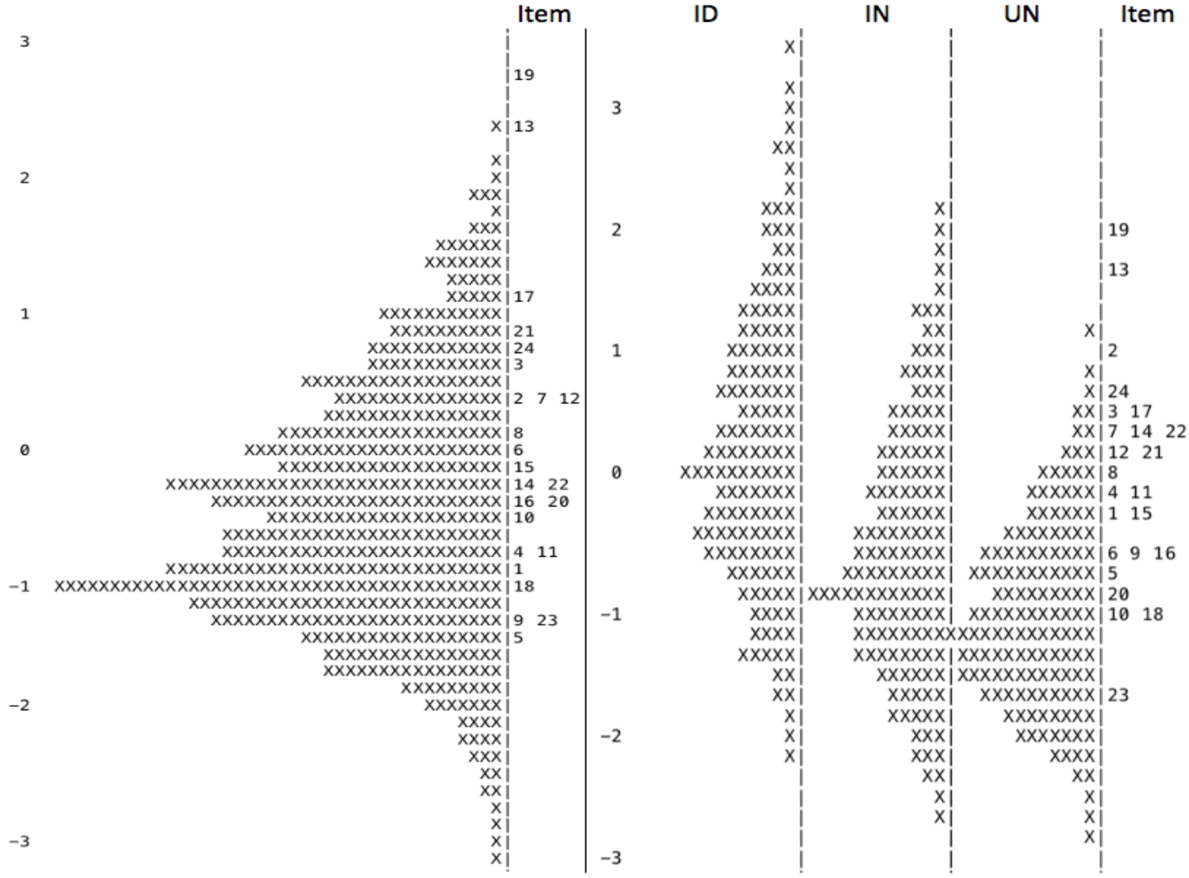
Table 6.4 shows the psychometric properties of the test item in unidimensional and three-dimensional models. The infit indices ranged from 0.84 to 1.27, and the average value was 0.99 digits (SD = 0.10) in unidimensional model, while three-dimensional model had an infit range from 0.85 to 1.23 and a mean infit of 1.0 (SD = 0.1), indicating that the test is a reliable construct and statistical fit. The results showed that item difficulty values were between −1.43 and 2.75 and −1.77 to 2.03 in unidimensional and three-dimension models, respectively.

Table 6.4. Summary of psychometric characteristics of test items.

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

Note. Dis: Discrimination; Dif: item difficulty.

Figure 6.2 illuminates the Wright map presenting the relationship between individual responses and item difficulty estimates. It seems that most of the items fall into the middle score zones of the map. The item mean default was 0.00, and the person mean was located at −0.08 logits in the unidimensional model, indicating a little low competency among the participants. Regarding each subskill in the three-dimensional model, the average person proficiency of +0.70, −0.18, and −0.77 for ID, IN and UN, respectively. Compared to 0.00 digits of average item difficulty as defaulted in the program's setting, the students seem to be more proficient in the ID items but lower on IN and UN. They felt that the UN items were the most difficult. Most of the test items showed a good model fit to the empirical data, suggesting no redundancy or overlap in the items. The most difficult items were items 19 (UN07) and 13 (UN03), which scored 2.45 above the item mean, signifying that the students had difficulties solving these items. This also suggests that respondents seemed to have misconceptions in the indeterminacy of confounded experiments.



a) Unidimensional model, x = 1.5 cases.

b) Three-dimensional model, x = 5.6 cases.

Note. ID: Identifying controlled experiments, IN: Interpreting controlled experiments, UN: Understanding the indeterminacy of confounded experiments.

Figure 6.2. Wright map of person-item location distribution in the test.

6.4.2 Internal consistency estimate and correlations

The full test results showed an MLE person reliability of.80 (Cronbach’s alpha = .81; McDonald’s omega = .82), Expected - a posteriori/plausible value reliability of .78, and item reliability of .99. Table 6.5 presents the internal consistency estimates and intercorrelations for each subskill. For our study’s purposes, these were acceptable levels in terms of internal consistency reliability in general (Taber, 2018).

There was a significantly positive correlation between the three subskills of CVS, with Pearson’s correlations ranging from 0.480 to 0.611. The results were consistent and even somewhat higher than these coefficients in the recent study by Schwichow et al. (2020), which ranged from 0.33 to 0.56.

Table 6.5. Internal consistency estimates and intercorrelations (Pearson) between subskills.

	α	ω	EAP/PV	Identifying	Interpreting
Identifying	.66	.70	.79		
Interpreting	.65	.68	.79	.611***	
Understanding	.55	.57	.77	.480***	.540***

Note. α: Cronbach’s alpha; ω: McDonald’s omega; EAP/PV: Expected - a posteriori/plausible value.

****p* < .001.

6.4.3 Impacts of subskills, content and number of independent variables on item difficulty

As suggested in the literature, the item difficulty was investigated in incorporating subskills, physics-related content and a number of changeable variables in the experimental set. Figure 6.3 depicts the item difficulty with respect to subskills of CVS. The items belonging to the understanding task were most difficult (M = 0.618, SD = 1.430), followed by items in the interpreting (M = 0.071, SD = 0.580) and identifying tasks (M = −0.690, SD = 0.570). One-sample Kolmogorov-Smirnov test confirmed that this item difficulty data was normally distributed (D = 0.122, *p* = .864). The ANOVA analysis compared the effect of these subskills on item difficulty. The results indicated a significant difference in item difficulty among subskills tasks (F (2, 21) = 3.823, *p* = .038). Tukey’s HSD test was implemented as a *post hoc* analysis to investigate differences in pairs of subskills tasks by making multiple comparisons. A significant difference was indicated between understanding and identifying items (Cohen’s d = 1.200, *p* = .031), but no significant difference was found between the understanding and interpreting items (*p* = .267) or between the interpreting and identifying items (*p* = .494).

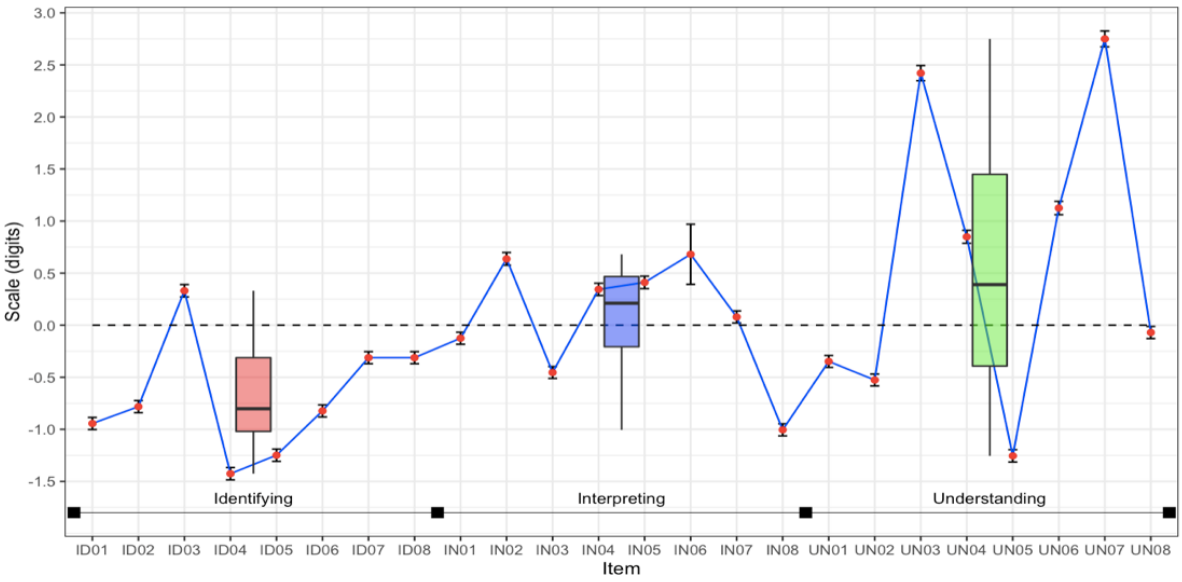


Figure 6.3. The item difficulty and boxplots of items following subskills.

Figure 6.4 illustrates the item difficulties regarding content knowledge. We examined the impact of the physics-related content on the item difficulty in the test. The mean of item difficulties of items belonging to electricity and electromagnetism (M = −0.2656, SD = 0.849) were lower than those belonging to mechanics (M = 0.137, SD = 1.153), and heat and thermodynamics (M = 0.067, SD = 1.266). Nevertheless, an ANOVA analysis for the three traits was employed to detect whether the item difficulty depends on the content. The result revealed that the content knowledge did not significantly affect item difficulty in this study (F(2, 21) = 0.334, *p* = .793).

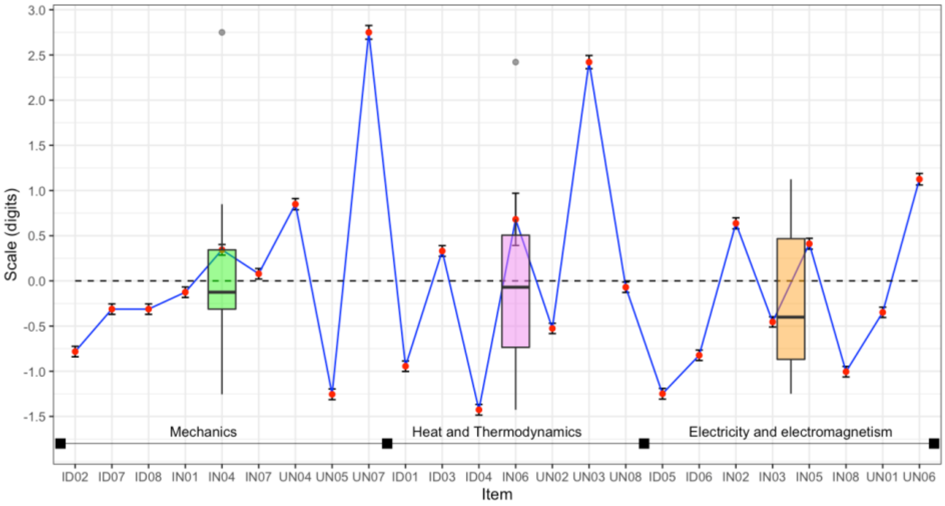


Figure 6.4. Item difficulty and boxplots of item bundles regarding physics-related content.

Figure 6.5 demonstrates the pattern of impact of the proportion of independent variables on item difficulty. Surprisingly, the item difficulties of items with three variables seemed to be lower than those with fewer variables. However, we applied an ANOVA analysis to examine whether the number of independent variables in an experimental system influenced the item difficulty. The results showed that the proportion of independent variables did not significantly impact item difficulty in F (2, 21) = 1.256, *p* = .342.

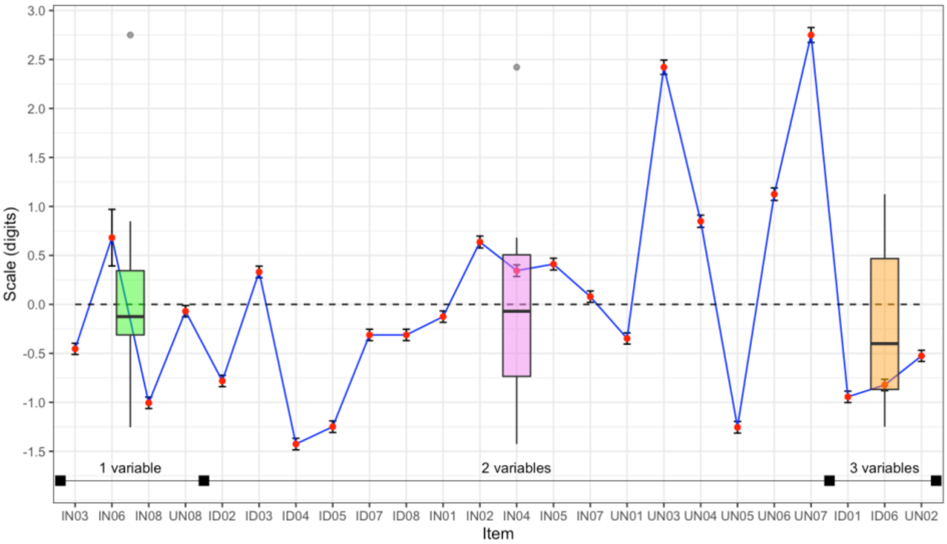


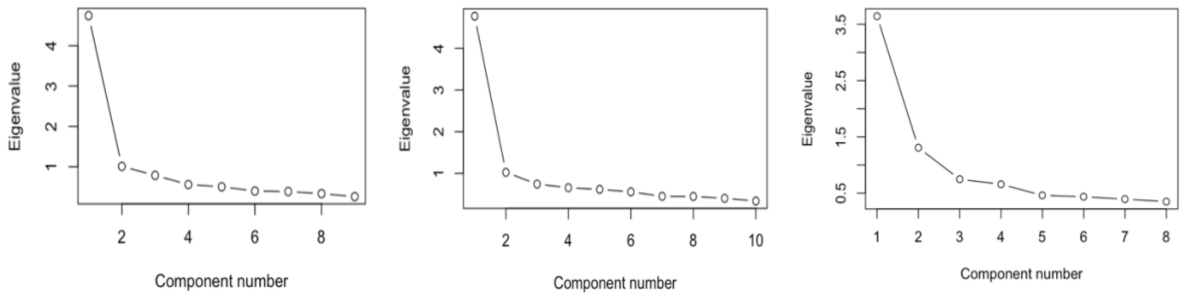
Figure 6.5. The item difficulty and boxplots of item bundles in regard to the number of independent variables.

6.4.4 Validity and reliability of the adapted students' motivation in learning physics questionnaire

6.4.4.1 Principal component analysis

The primary purpose was to identify and calculate scores for the factors underlying each scale and as suggested by former study (Martin et al., 2016), we implemented PCA as a prior condition for further analysis. As presented in initial eigenvalues and scree plot in Figure 6.6, the first principal component explained 59%, 57% and 50% of the variance in the LL, ET and CL scales, respectively. The second principal component had eigenvalues still over one, but each just explained around 15% of the variance in each scale. We employed analyses to solve with one factor and two factors by using varimax rotations of the factor loading matrix. The first principal component for each scale was considered due to three reasons: (1) it was supported by the former studies (Hooper et al., 2013); (2) the eigenvalues on the scree plots tended to level off after two components; and (3) the loading indices in one-principal component models were higher than 0.30 each. Particularly, the component loadings of each questionnaire item were positive and sizable, indicating a strong correlation between each item and the scale. The results of varimax rotation showed standardized loadings ranged from 0.62 to 0.83 for the LL scale, from 0.46 to 0.77 for the ET scale and 0.64 to 0.72 for the CL scale.

Generally, this suggested that all items should be kept, and the items in each scale were adequate as a single factor scale.



a) Like learning b) Engaged by teaching c) Confidence in learning

Figure 6.6. The scree plot in PCA for each scale.

Reliability coefficients of Cronbach's alpha for each one-principal component scale were calculated in the case. These values were .90, .88 and .86 for the LL, ET and CL scales, respectively. Overall, it is very good in terms of internal consistency reliability at all scales (Taber, 2018).

6.4.4.2 Rasch model measurement analysis

The results of the PCM (Adams & Wu, 2010) in Rasch measurement revealed that the scales fit quite well models to the data. As designed as unidimensional models (Martin et al., 2016), three scales performed consistently in one-dimensional structure.

Generally, all items in the three scales fit well to the present dataset. For the LL scale, the infit indices ranged from 0.81 to 1.24, and the location parameter (item difficulty) values were in the range of −0.729 and 0.942 (M = 0.00, SD = 1.00). All infit indices met very well to the cut-off standards for the ET scale with the infit values ranged from 0.84 to 1.30 (M = 0.99, SD = 0.15) and CO scale with infit indices ranged from 0.98 to 1.06 (M = 1.02, SD = 0.03). The item location parameters in the ET scale and CO scale ranged from −0.362 to 0.699 and from −0.613 to 1.149, respectively. The means of person response were 1.184 digits (SD = 1.965), 1.356 (SD = 1.973) and −0.351 (SD = 1.727), indicating that students had positive attitudes to like learning physics and felt that they received much engagement from their teacher in learning physics. However, students reported lowest scores in the confidence in the learning physics scale.

Additionally, we implemented DIF analysis with respect to gender and grade levels by using ordinal regression methods in R lordif package (Choi et al., 2011) (see Section 3.3.3). The likelihood ratio χ^2test is considered as the detection criterion at the α level of 0.01, while the change in McFadden's R^2 of 0.02 as a criterion for rejecting the null hypothesis of no DIF. All pseudo R2 values were below 0.02, and p-values of the goodness-of-fit statistics above 0.05, indicating that no DIF items were detected in all three scales regarding gender.

6.4.4.3 Internal consistency and correlations

Table 6.6 presents the internal consistency estimates and intercorrelations for each scale. Internal consistency reliability was an adequate level for all single scales (Taber, 2018). A significantly positive correlation was found between the three scales of physics motivation. The strongest association was found between LL and CL scales, followed by LL and ET scales.

Table 6.6. Internal consistency estimates and correlations between subskills.

	α	ω	EAP/PV	LL	ET
LL	.89	.91	.87		
ET	.87	.89	.86	.604***	
CL	.83	.88	.83	.646***	.407***

Note. α: Cronbach's alpha; ω: McDonald's omega; EAP/PV: Expected - a posteriori/plausible value; ****p* < 0.001.

6.4.5 Different performance among grade cohorts on the CVSP test

Figure 6.7 illustrates students' achievements on the CVSP test across grade levels. There were numerous students who received a very low score in the 8th (M = −1.053, SD = 0.837) and 9th grade (M = −0.946, SD = 0.902), and some 9th graders gained very high scores. Distribution in the 10th graders' scores was similar to those in the 8th and 9th grades, but it scored somewhat higher on average (M = −0.443, SD = 0.992). The mean scores of the 11th- and 12th-grade groups were 0.147 and 0.560 digits (MLE), respectively, but their scores' distribution showed in different patterns. The 11th-grade cohort had both the high- and low-performing participants. It appeared that there was an equivalent proportion of students who achieved higher mean scores and lower scores in this group. In the 12th grade cohort, there were still some students who achieved the highest scores on the test, so its distribution shape tended to spread up the top of the scale.

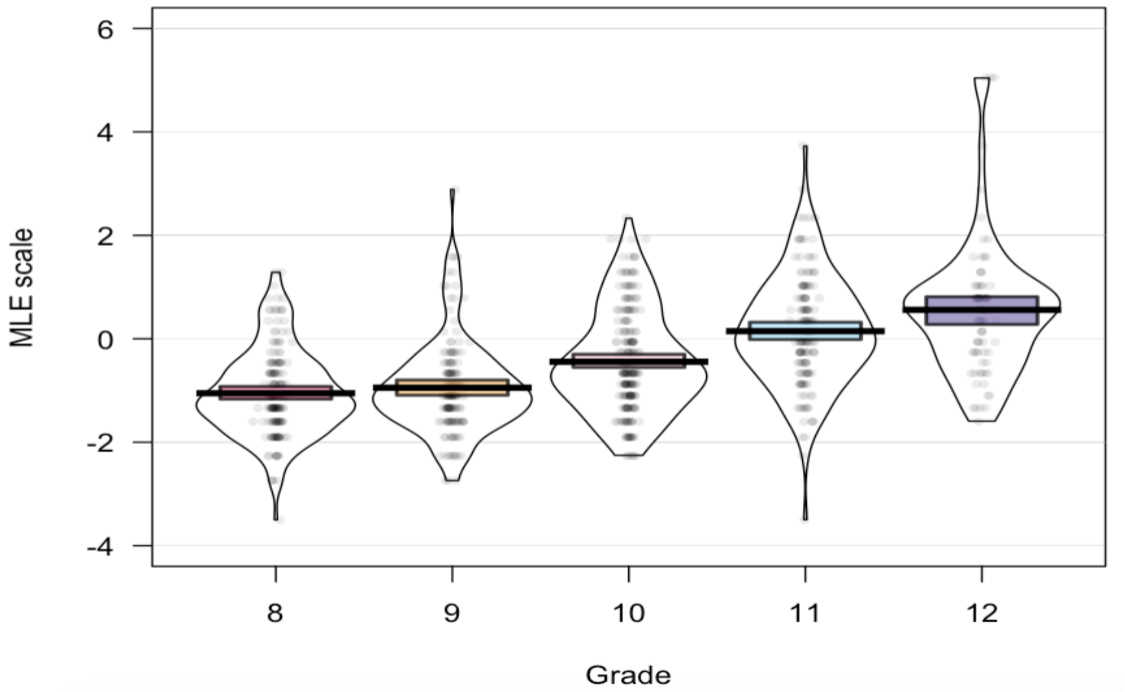


Figure 6.7. Differences in performance of grade cohorts on the CVS test.

Furthermore, we applied the four-parameter symmetric log-logistic equation (Kniss AR, 2018) (see Section 3.3.4) to describe the developmental process in students' CVS capacities. The parameters of a sigmoid curve were tailored by the R drc package (Ritz et al., 2015) and depicted via smooth line graphs in the R ggplot2 (Wickham, 2016). As shown in Figure 6.8, the fitted logistic curve described the empirical data adequately in comparison with the log-logistic model. Result of comparison in the dose-response model to a more general ANOVA model which employed an approximate F-test (Kniss AR, 2018), revealed that the logistic curve fit quite well to the empirical data (F = 0.332, p = .565). In general, development of children's CVS in learning physics was statistically significant across grade cohorts, but growth rates were different between particular cohorts. The most rapid growth was flagged at the 10th grade as the point of inflexion (ED50 - half maximal effective concentration) (See Section 3.3.4). This suggested that the fastest development occurred from the 9th grade to 11th grade ($tan\mathrm{M}_{\mathrm{ED50}}$ = 2.978).

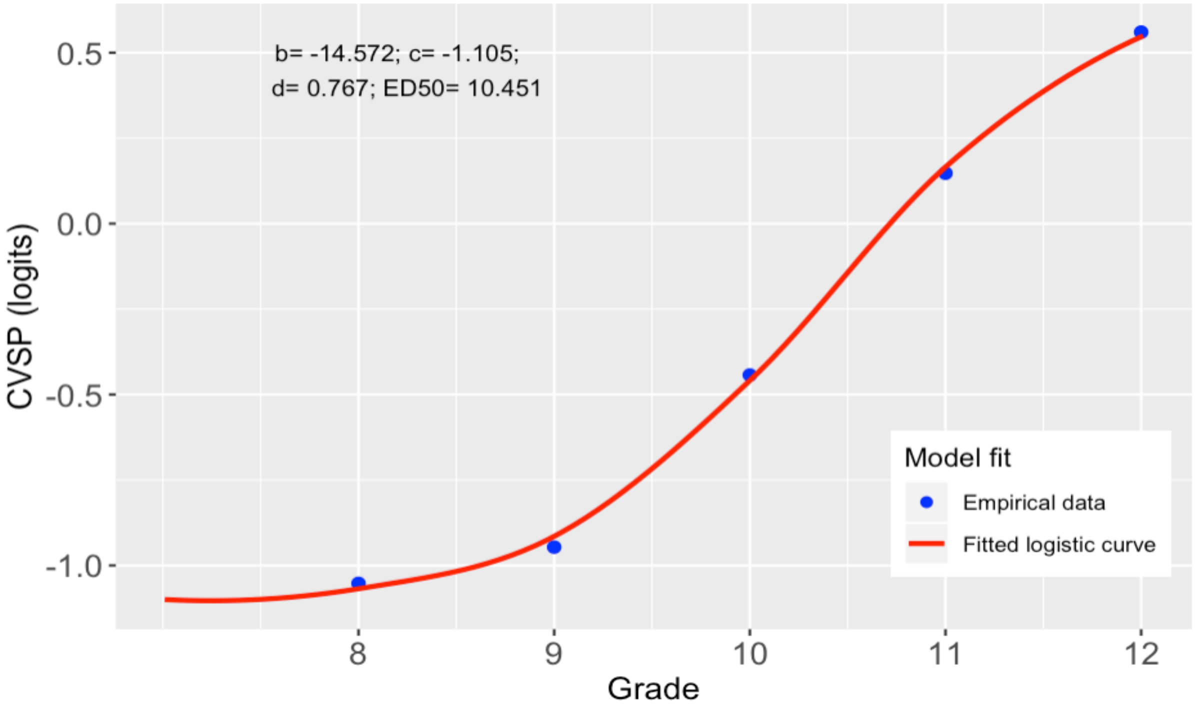


Figure 6.8. Developmental curve of CVSP in physics.

6.4.6 Gender difference in CVS

As demonstrated in Figure 6.9, males and females seemed equivalent in CVS ability at each cohort and whole sample. In the 8th grade to 10th grade, the average score of males were a little higher that of females, and the distributions displayed slightly differently regarding gender, with the larger part of the density of the male spread out over a wider range. This verified that the standard deviation was larger in the male cohort. However, males appeared to earn a higher mean score than females.

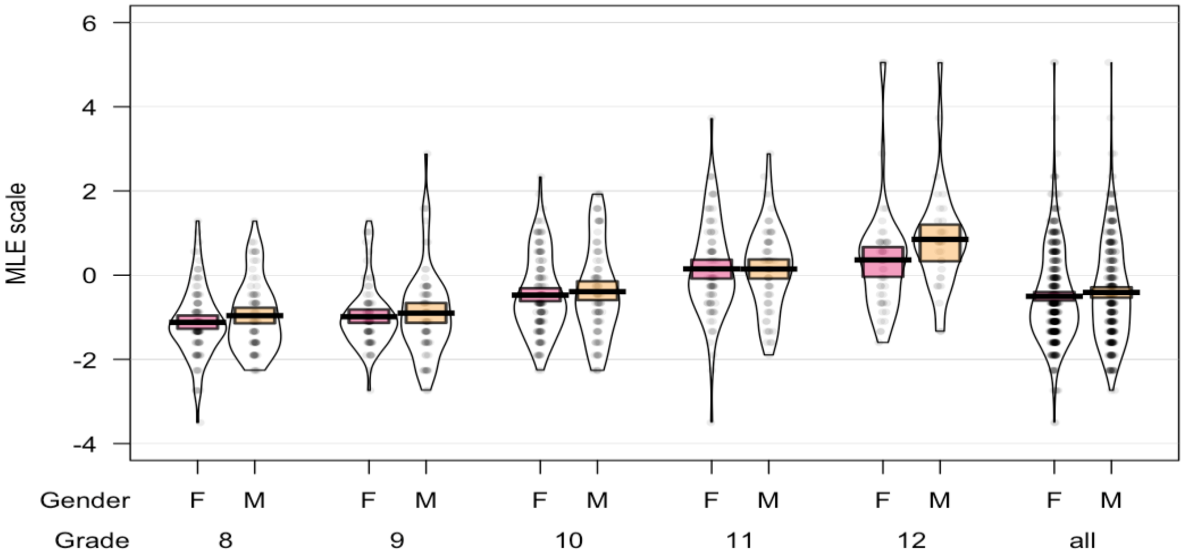


Figure 6.9. Comparison of performance in CVS in physics between males and females.

Furthermore, we employed the *t*-test to compare abilities between males and females. Table 6.7 summarized the results of the *t*-test for five cohorts and the whole sample. No significant difference was found on the CVSP test between boys and girls in the whole sample or single cohorts. This also suggested that boys did not differ significantly from girls on the CVSP test. In other words, it was estimated that males and females had an equivalent ability level on the CVSP test. Moreover, Table 6.8 provides the results of the *t*-test concerning gender in individual subskills. The results indicated that no gender difference was found on the ID and IN tasks, but male students performed significantly better than female students on the UN task.

Table 6.7. The *t*-test to compare CVSP test results regarding gender.

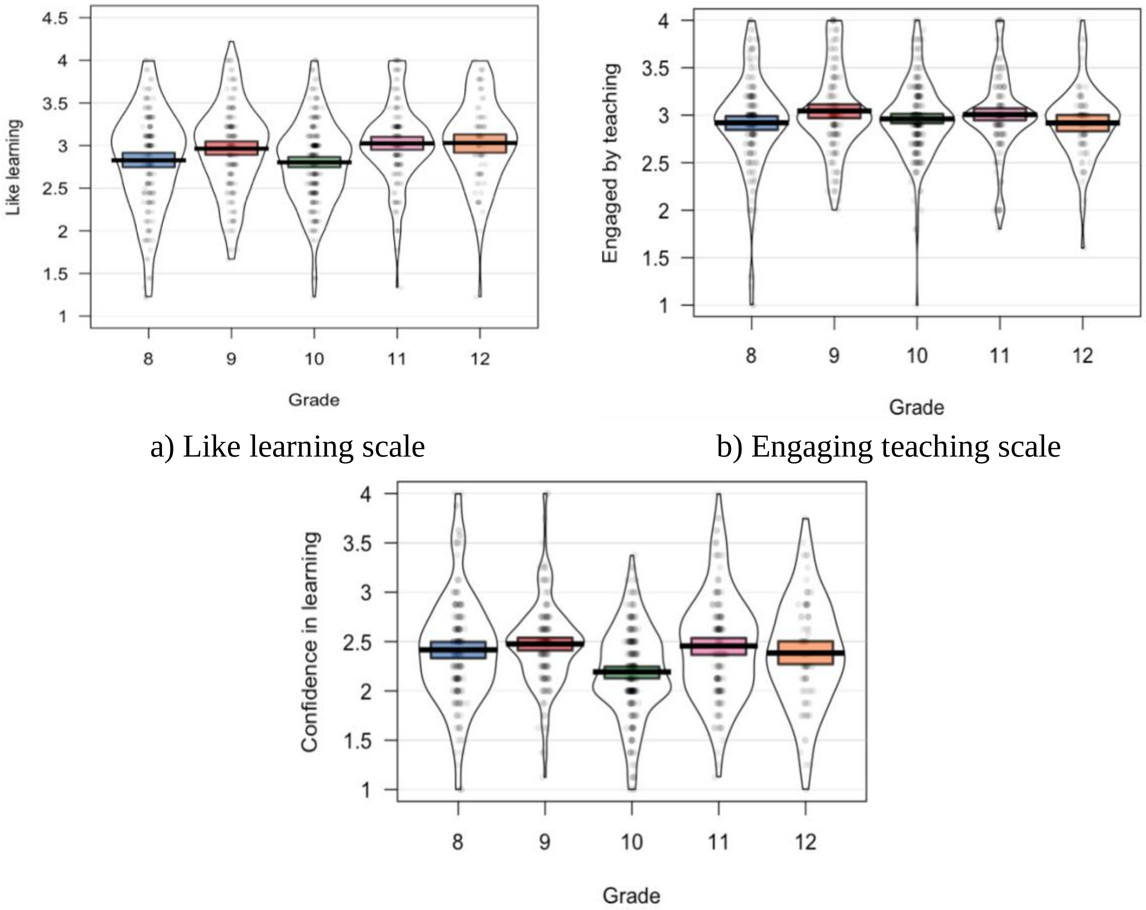
	Male		Female			
Grade	N	*Mean (SD)*	N	*Mean (SD)*	t	p
8	77	−0.96 (0.85)	101	−0.12 (0.83)	1.25	.212
9	77	−0.90 (1.04)	82	−0.98 (0.75)	0.58	.560
10	99	−0.39 (1.09)	144	−0.48 (0.92)	0.60	.550
11	67	0.15 (0.99)	87	0.15 (1.13)	−0.02	.987
12	33	0.85 (1.25)	48	0.36 (1.32)	1.68	.097
All	345	−0.41 (1.75)	462	−0.50 (1.10)	1.13	.256

Table 6.8. The t-test to compare CVS abilities between males and females on each subtest.

Subtest	Male	Female		
	Mean (SD)	*Mean (SD)*	t	p
Identifying	0.26 (1.54)	0.22 (1.45)	0.41	.683
Interpreting	−0.56 (1.50)	−0.65 (1.43)	0.85	.397
Understanding	−0.97 (1.30)	−1.18 (1.21)	2.41	.016

6.4.7 Physics motivation in students at different grade cohorts

The pirate plots in Figure 6.10 illustrate the students' performance on the (a) LL, (b) ET and (c) CL scales across five grade cohorts. Overall, perspectives toward learning physics of children were positive and quite high. The mean scores seemed to be similar among five cohorts, except at the 10th grade, where students scored lower on the LL and CO scales. On average, students scored around 3 points (in 4-point Likert scale) on the LL and ET scales, and just under 2.5 points on CL scale. The changing trend seemed to be the same pattern in individual motivational factors across grade levels.



a) Like learning scale

b) Engaging teaching scale

c) Confidence in learning scale

Figure 6.10. Performance of students in different grade levels on each scale in physics motivation.

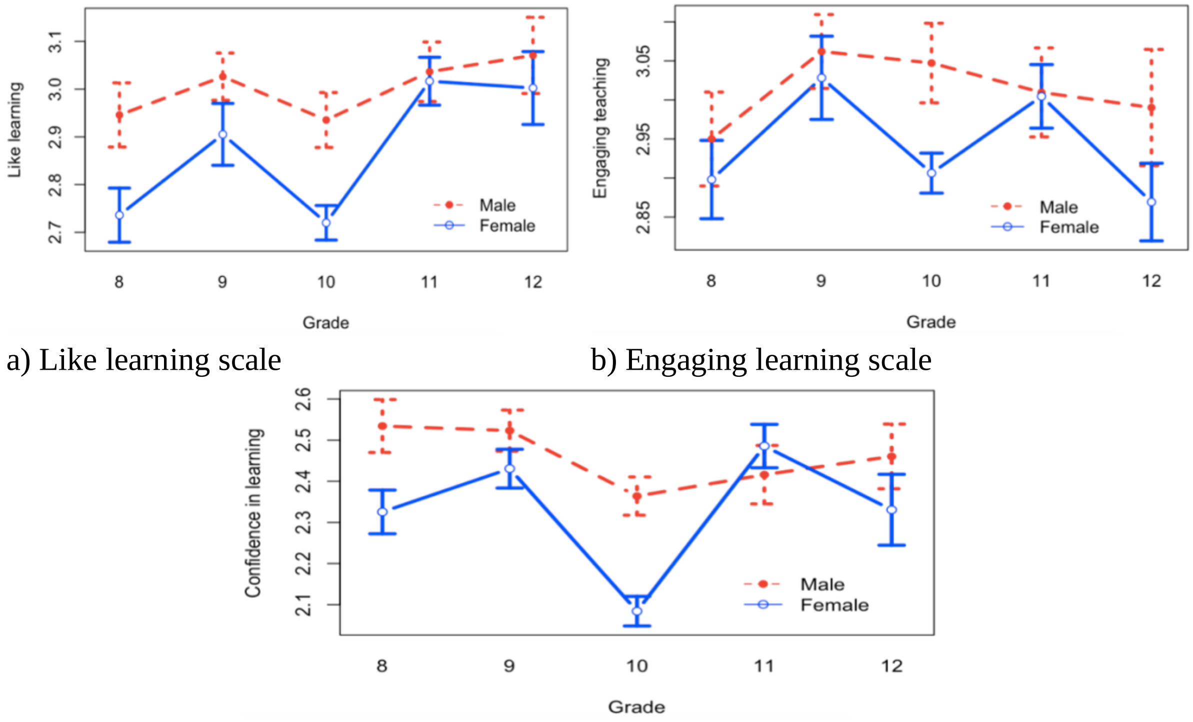
Moreover, the ANOVA was applied to investigate the effect of grade levels on students' motivation toward learning physics. The results indicated that there was a non-significant difference between grade levels on the ET scale [F(4, 802) = 2.3, p = .057]. By contrast, a statistically significant difference between the grade cohorts was found on the LL [F(4, 802) = 7.0, p < .001] and CL scales [F(4, 802) = 10.7, p < .001]. Tukey's HSD was employed to test significant differences in pairs of grade levels. Table 6.9 summarises the multiple comparisons, the younger groups tended to gain higher scores than the older ones on the LL scale. The 10th graders showed significantly lower scores than those of the 11th and 12th graders on both these scales.

Table 6.9. Tukey's HSD in multiple comparisons on the LL and CL scales.

Grade	Like learning		Confidence in learning	
	Mean difference	p	Mean difference	p
8 & 10	0.02	.991	0.22	<.001
8 & 11	−0.20	.005	−0.04	.952
8 & 12	−0.20	.029	0.03	.989
9 & 8	0.14	.112	0.06	.809
9 & 10	0.16	.023	0.28	<.000
9 & 11	−0.06	.833	0.02	.997
9 & 12	−0.07	.883	0.09	.661
10 & 11	−0.22	<.001	−0.26	<.001
10 & 12	−0.23	.007	−0.19	.026
11 & 12	0.00	1.00	0.07	.833

6.4.8 Gender difference in physics motivation

Figure 6.11 shows performance of males and females on each physics motivation scale across grade levels. Although a fluctuation in motivation can be observed on the single scales across grade cohorts, a general tendency was that boys reported higher than girls on most scales, except in the 11th grade, where girls achieved higher scores than boys did on the CL scale.



a) Like learning scale

b) Engaging learning scale

c) Confidence in learning scale

Figure 6.11. Comparison of students' performance regarding gender across grade levels.

Furthermore, the *t*-test was manipulated to compare students' motivation toward learning physics with respect to gender. For the entire sample, a significant difference was determined between males and females on all three scales, with favouring males (Table 6.10). In fact, the significant difference was only found at the 8th and 10th grades. No significant difference with respect to gender was found in other grade cohorts, although boys mostly scored higher than girls.

Table 6.10. Comparing between males and females in motivation toward learning physics.

Grade	Like learning			Engaging teaching			Confidence in learning		
	ΔM	df	d	ΔM	df	d	ΔM	df	d
8	0.21	160.7	0.38*	0.05	159.8	0.09	0.21	158.8	0.35*
9	0.15	148.9	0.20	0.03	155.8	0.03	0.09	155.8	0.21
10	0.21	159.7	0.51**	0.14	135.6	040*	0.28	187.8	0.58***
11	0.02	135.5	0.05	0.00	125.4	0.05	-0.07	128.8	0.16
12	0.07	74.75	0.11	0.12	58.9	0.37	-0.13	78.2	0.27
All	0.14	745.3	0.29***	0.08	680.1	0.18*	0.16	744.6	0.34***

Note. ΔM: mean difference between boys and girls. d: Cohens' d effect size.
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

6.4.9 Interaction among grade level, motivation and physics test in predicting CVS

A path analysis was yielded with Mplus 7 (Muthén & Muthén, 2012) to explore underlying factors explaining students' CVS capacity. As suggested in the literature, we assumed exploratory variables such grade level (GR), physics test's result in previous semester (PH), like learning scale (LL), engaging learning scale (ET) and confidence in learning (CL) as the interactive factors in predicting individuals' CVS abilities.

As illustrated in Figure 6.12, the model presents relationship between understudied variables in contributing to students' CVS proficiency in physics. The results revealed that the model fit very well to the empirical data, with χ^2(11) = 23.16, p = .017, CFI = .994, TLI = .984, SRMR = .021, RMSEA = .037. The model enabled to explain around 44.7% of variance of the CVS capabilities. GR demonstrated as the strongest factor affecting directly on the CVS, followed by PH and LL variables. However, ET and CL do not contribute directly to explaining significantly on the CVS, but CL showed as the main predictor of PH.

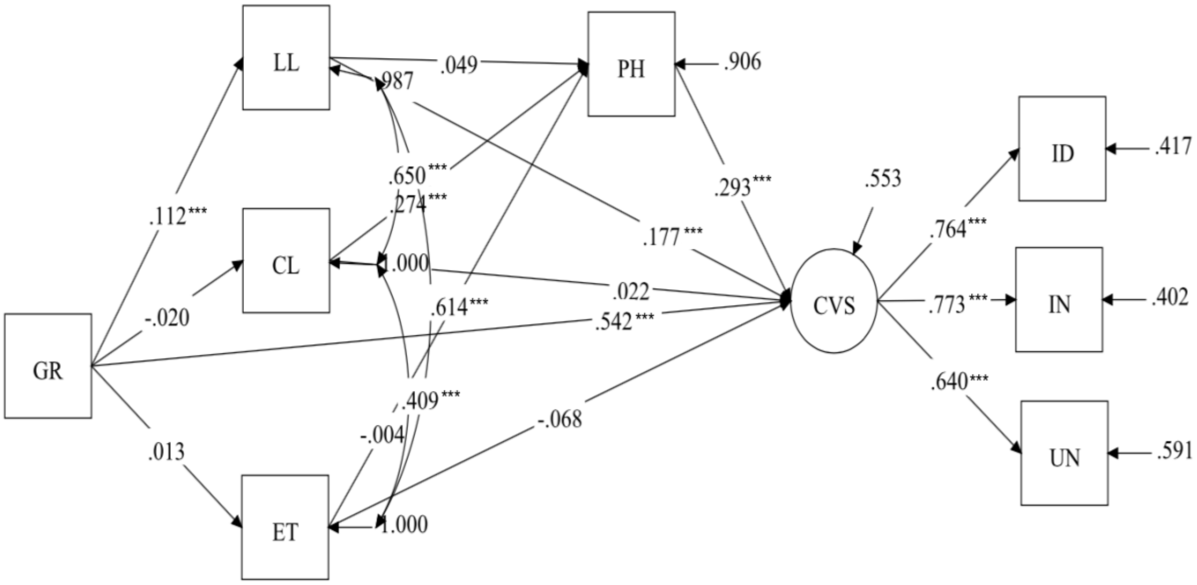


Figure 6.12. Proposed model predicting students' CVS abilities.

Note. GR: grade level; PH: physics test in previous semester; LL: like learning physics scale; ET: engaging learning in physics; CL: confidence in learning physics.
***p<.001.

6.5 Discussion and conclusion

The subskills of CVS, such as planning the suitable experimental structure, identifying the appropriate variables, interpreting the valid outcomes and understanding the contribution of components in an experimental system, are the core scientific reasoning capacities in learning physics. As Chen & Klahr (1999) argued, it is very important to know how well present school practices have affected CVS in school contexts. The current study strived to assess students' CVS skills at different grade levels. Some validity evidence for the proposed test was provided, involving content validity for using robust statistical measures to explore the psychometric properties of the test with the Rasch model and DIF analysis regarding gender. This supports the hypothesis H3.2. It seems that the CVSP test is well-targeted to assess three subskills of CVS in physics for secondary school students.

Considering the test dimensionality, interestingly, both the unidimensional model and three-dimensional model performed well since both fit quite well to the current dataset. The hypothesis H3.1 is confirmed. The results can be interpreted in terms of single subskills of CVS contributing to a common underlying one-factor structure. Rasch model analysis suggested the well-infit individual items in the test, but classical statistics recommended that a few items need to be reconsidered for future improvement. According to the classification of Ebel and Frisbie (1991), item 23 (UN05) is a poor item since its discrimination value was 0.08

(lower than 2.0), and four items (ID03, IN02, UN03, UN07) had a discrimination range between 2.0 and 2.9, prompting minor revision.

Furthermore, subskills' impacts on item difficulty were illuminated in the Wright persons-items map of the three-dimensional model and further confirmed statistically by ANOVA analysis. The cluster of ID items was easier than others, while students felt most challenged by the items related to their understanding of confounded experiments' indeterminacy. Amazingly, although this study included students from five grade levels, physics-related content knowledge did not significantly influence the item difficulty. These results are consistent with the literature, which revealed the subskills impacted item difficulty, but the content knowledge did not (Schwichow, Christoph, et al., 2016). The findings confirm the hypothesis H3.3. Nevertheless, their study showed that significant differences were detected both between UN and ID and UN and IN items. In contrast, the present study only found a significant difference between UN and ID items with a smaller effect size. Adding the independent variables into an experimental system might make a problem more difficult because of the higher working memory load. However, the current study revealed no significant difference among one to three variables in terms of item difficulty. Not surprisingly, these results are consistent with the study by Staver (1986), which indicated that the effects of the number of independent variables on item difficulty only happened significantly in the four- and five-variable items.

Regarding grade level, CVS capacity increased across grade cohorts. The findings are apparently consistent with the findings of former studies (e.g. Han, 2013; Schwichow et al., 2020; Zimmerman, 2007). Moreover, the developmental curve was modelled within a four-parameter symmetric log-logistic equation. The model performed as a good fit to the empirical data, indicating that the fastest improvement of CVS happened at the 10th grade (15 - 16 years). The results support the hypothesis H3.5. However, a non-significant difference on the ET scale, but a significant difference between the grade cohorts was found on the LL and CL scales. The hypothesis H3.6 is partially confirmed.

Generally, DIF analysis confirmed that the test and questionnaire items were equivalent, because no DIF was flagged between boys and girls statistically. However, when focusing on each subskill of CVS with *t*-test, we found that males did not differ significantly to females in the ID and IN subskills, but the UN items seemed to favour the males since scoring significantly higher than did males. The results were partly in agreement with the previous (Mayer et al., 2014; Piraksa et al., 2014). For physics motivation, the psychometric properties of the adapted questionnaire were examined in Vietnam contexts. The results confirm the hypothesis H3.4.

The findings showed that there was a significant difference in all motivation scales, in which males reported higher scores than females. The hypothesis H3.7 is not confirmed in the case.

Investigating the relationships between CVS and relevant variables (e.g., physics test, GPA) and examining the underlying factors predicting the item difficulty are additional evidence of validity arguments. Most of the findings were consistent with the former studies in the field (The hypothesis H3.8 is supported). Notably, the study by Valanides (1997) indicated that achievement in mathematics and GPA contributed considerably to predicting SR involving CVS tasks, while Schwichow et al. (2020) found that school grade and content knowledge tests in physics were significantly associated with each subskill of CVS. The path analysis also confirmed the important role of content knowledge and physics motivation in predicting CVS ability in children.

6.6 Implications in science education

The study has partly contributed the potential items to the field of assessment of CVS reasoning ability. It also provided the highlighted construct of problem items that teachers can refer to when developing the CVS tasks. Teachers can apply appropriate tasks relevant to students' competencies in classroom settings and labs as hands-on activities. Although males did not differ statistically from females in CVS tasks generally, there was a slight difference in each subskill. Consequently, teachers should pay attention during instruction to help male students improve in ID tasks and support female students deal with UN tasks more effectively.

Investigating the impact of these subskills, knowledge content and the number of independent variables on the item difficulty plays a notable role in school practices. The main challenge of an item is not to exclude knowledge concepts (abstract, complex, and difficult) or the number of independent variables in an experiment, but it comes from the kind of subskills of the CVS task. The study also revealed that students developed CVS with the fastest rate during the first year at high schools. Teachers need to be aware of this tendency to select suitable activities to enhance students' CVS skills. Consistent with the previous studies, students seem not to distinguish between the confounded experiments and controlled experiments. As Siler and Klahr (2012) discussed, a possible reason for this inclination may derive from procedural misconceptions. Students often misunderstand the goal of the task as contriving an outcome rather than finding out about the causal status of a single variable. Students believe that they can "do it all" in a single experiment. Still, if two variables change simultaneously, we can examine the effect of each variable on the outcome of an experimental system. Therefore, teachers should help students be aware of these possible pitfalls in executing the CVS tasks.

Nonetheless, the current test did not consider assessing the planning-controlled experiments skills of students. In principle, developing this kind of problem in the virtual environment is possible, using simulated real-life experimental tasks in advance (Luecht & Sireci, 2011). The efforts need to be encouraged for the next generation of the test. Additionally, no items provide experimental data, so we could not measure the interpreting skill of students that require a quantitative dataset. According to Han (2013) and Zhou et al. (2016), CVS tasks with quantitative data can tap into students' thinking beyond the testability of the variables that influence the results of an experiment. This kind of stem of items needs to be considered for future research. Furthermore, the text length and use of visual images may affect the item difficulty (Stiller et al., 2016), but the current study did not handle the issues. Future studies would deem the impacts of multiple representations on item difficulty.

CHAPTER 7. EFFECTS OF MULTIMEDIA ON PSYCHOMETRIC PROPERTIES OF THE COGNITIVE TESTS: A COMPARISON BETWEEN TECHNOLOGY-BASED AND PAPER-BASED ASSESSMENT

7.1 Introduction

ICT has become increasingly relevant in most aspects of modern life, including work and school. Computers have been important supplementary tools in the teaching and learning process. As part of this explosion of new technologies, TBA is being used globally on either computers or other electronic devices, such as smartphones, tablets and other portable devices. TBA provides evidence of positive results on student learning performance, motivation and attitudes (Mohamadi, 2018; Nikou & Economides, 2018; Sheard & Chambers, 2014). TBA offers numerous benefits, such as a higher standardization of test administration, efficient test scoring, and the likelihood of immediate reporting and interpreting of results (Csapó et al., 2012; Shute & Rahimi, 2017). A large number of institutions have considered introducing TBA administration in recent years, and thus TBA has progressively replaced traditional paper-and-pencil assessment.

Creativity, critical thinking, problem-solving and ICT are all regarded as significant aptitudes in the 21st century (Voogt & Roblin, 2012). Moreover, students are expected to demonstrate these abilities and skills to meet the demands of future jobs. IR and scientific reasoning in the CVS are closely tied to problem-solving and play an important role in learning core school disciplines (e.g., Adey & Csapó, 2012; Chen & Klahr, 1999; Van Vo & Csapó, 2021a, 2021b). Modern society is under ever greater pressure to deal with more information in a shorter amount of time. Assessment of these skills has therefore come under increasing consideration. Like other testing in educational contexts, these cognitive tests can be used in testing mechanisms in a virtual environment, so they have gradually transformed from traditional forms into technology-rich formats.

The paper method is restricted to using static text and graphics, whereas computers are capable of presenting rich visualizations of figures and even dynamic interactions with test-takers (Greiff et al., 2018). However, previous studies have reported inconsistent findings when comparing performance between dual modes of administration (Gates & Kochan, 2015; Williamson et al., 2017). There are various approaches to investigating equivalent groups in these studies, such as construct validity and mean scores from test to item levels (Gates & Kochan, 2015; Williamson et al., 2017).

In the context of Vietnam, most schools were at the *infusing* stages of ICT-facilitated teaching and learning pedagogies, and the conditions in schools were suitable for the necessary transformation of their ICT-facilitated teaching and learning practices (Maftuh, 2011). Nevertheless, the development of assessment was thought to be at the *applying* stage, meaning that teachers had started learning to use ICT along with traditional methods. In recent years, schools have been interested in using technology in educational assessment, although paper-and-pencil mode is still used officially in Vietnam. Institutions often offer both paper and TBA administration because of the inadequacies of the current ICT infrastructure. Specifically, when novel coronavirus (COVID-19) broke out, it posed a major challenge for the traditional school environments. Testing administration in education should be in line with the national restrictions on social interactions due to the pandemic. Therefore, studies comparing the impact of the paper - based and TBA modes of administration warrant a scholarly inquiry. The main aim of our study is to investigate the effects of technology in different modalities in cases with and without supervision on cognitive reasoning tests.

7.2 Research questions

This study aimed to evaluate the extent to which the administration modes affect the cognitive tests. We are interested in the differences of validity of the internal structure and scores when comparing the dual delivery modes on the same test, taking into consideration the item, task and test levels. The following three research questions guided our study:

RQ4.1: Are the adapted tests suitable for students in the Vietnamese context using the two methods of delivery?

H4.1: We expect the tests are equivalent in term of internal structure validity in two administration modes (e.g. Hassler Hallstedt & Ghaderi, 2018; Kim & Huynh, 2010; Neumann & Neumann, 2019; Schroeders & Wilhelm, 2010)

RQ4.2: Is there any evidence of equivalence between the online and PP groups in the cognitive tests at item and task levels?

H4.2: We hypothesize that it is comparable for two groups' performance at the item and the bundle of items (Kim & Huynh, 2010).

RQ4.3: How do the different administration modes affect students' performance at test level?

H4.3: It is assumed that no statistically significant difference in participants who are the same prior proficiencies on the test.

7.3 Data analyses

We involved three main analytical approaches at the item, task and test levels to evaluate the performance comparability in the two administration means. First, raw scores of each item were scored 1 and 0 for correct and incorrect answers, respectively. We then scaled the test results with the Rasch model, which is based on IRT to examine the relationship between the test-takers' abilities and their responses to the test items. Three main indices (final deviance, AIC and BIC) were referred to compare model fit in Rasch model (see Section 3.3.2).

Furthermore, DIF analysis provides evidence about functions to compare different groups at the item level. In the study, we employed DIF analysis in dichotomously scored items with the Angoff's delta plot method (see Angoff, 1982). DBF or item bundle DIF analysis is the natural extension of the DIF approach. In the current study, we applied the framework of Douglas, Roussos and Stout (1996), which used the SIBTEST method (Shealy & Stout, 1993) to examine DIF in bundles (see Section 3.3.3).

7.4 Paper-based and online testing in the IR test

7.4.1 Participants

The data set was drawn from 715 students in six public schools in the southern province of An Giang (Vietnam). A total of 20 classes were recruited for this study, based on matching the equivalence of students' school performance in the last semester from the sample cluster list of 20 public schools. Table 7.1 presents the basic characteristics of the four student cohorts of the 6th, 9th, 10th and 11th grades.

Table 7.1. Characteristics of the participants in study 1.

Grade	N	Online/PP ratio (%)	Online		PP	
			Male/Female ratio (%)	Mean age (years)	Male/Female ratio (%)	Mean age (years)
6	103	42.7/57.3	34.1/65.9	11.3	47.5/52.5	11.2
9	115	43.5/56.5	46.0/54.0	14.3	49.2/50.8	14.2
10	246	56.5/43.5	38.8/61.2	15.1	52.3/47.7	15.2
11	251	41.4/58.6	46.2/53.8	16.2	51.0/49.0	16.2
Total	715	47.1/52.9	41.5/58.5	14.8	50.5/49.5	14.8

Table 7.2 displays the equivalence of students' ability in terms of school performance during the previous semester by using Pearson's chi-squared test. School performance was computed as average of the 11–13 school subjects in both elective and compulsory subjects. In Vietnam secondary education level, this index is scaled into five categories: excellent, good,

fair, weak and poor. Students reported their school performance in the first part of the test. The distribution of school performance during the last semester showed the same between the online group and the paper-and-pencil group in the individual cohorts and entire sample.

We conducted this study in the first semester of the academic year 2019–2020. For the PP group, a test booklet with two single items on each page was given to each student along with an answer sheet. A teacher introduced the research aims and guided the students through the appropriate practice steps following our instructions. For the online test, each student received a link and an individual code to access the test. Two teachers were present to resolve technology issues during the testing process in the computer room. One item was displayed per screen page, and the items were presented in a manner similar to the PP version. Students spent around 30 minutes to complete this test and were supervised by their teachers as part of the regular school timetable. The online testing was administrated within eDia via the server of University of Szeged (Csapó & Molnár, 2019).

Table 7.2. Distribution of the online and PP groups based on school performance in the previous semester.

Grade	Mode		Poor	Weak	Fair	Good	Excellent	Total	χ^2	*p*
6	Online	N	0	0	1	19	24	44	1.387	.500
		%	0.00	0.00	2.27	43.18	54.55	100		
	PP	N	0	0	0	25	34	59		
		%	0.00	0.00	0.00	42.37	57.63	100		
9	Online	N	0	0	7	22	21	50	0.004	.998
		%	0.00	0.00	14.00	44.00	42.00	100		
	PP	N	0	0	9	29	27	65		
		%	0.00	0.00	13.85	44.62	41.54	100	0.240	.887
10	Online	N	0	0	2	65	72	139		
		%	0.00	0.00	1.44	46.76	51.80	100		
	PP	N	0	0	1	48	58	107		
		%	0.00	0.00	0.93	44.86	54.21	100		
11	Online	N	0	2	12	52	38	104	3.933	.269
		%	0.00	1.92	11.54	50.00	36.54	100		
	PP	N	0	0	12	74	61	147		
		%	0.00	0.00	8.16	50.34	41.50	100		
All	Online	N	0	2	22	158	155	337	2.492	.477
		%	0.00	0.59	6.53	46.88	45.99	100		
	PP	N	0	0	22	176	180	378		
		%	0.00	0.00	5.82	46.56	47.62	100		

7.4.2 Instruments

The original IR items were developed in Hungarian and included four tasks. This test has been used in several empirical studies in cross-cultural contexts to establish its reliability and predictive validity for use with school-age populations (e.g., Kambeyo & Wu, 2018; Korom et al., 2017). The basic criteria to select items were based on the structure of each item and the empirical evidence from the earlier studies. Ultimately, 20 items were used to measure students' IR capacity (see Section 3.2.1.1). The online test was developed using the eDia platform, an assessment platform developed by the University of Szeged. The eDia is a browser-based assessment platform. To ensure the same order of the items in both the online and PP versions, we designed a fixed-length test in the two versions in this study.

7.4.3 Findings of the study 1

7.4.3.1 Reliability and validity

Table 7.3 presents Cronbach's alpha (α), McDonald's omega (ω) and the Pearson correlation between the subtests. Generally, the internal consistency reliability was acceptable though not excellent for both test versions, but the PP format seemed somewhat better than the online one. There were significant positive correlations between the subtests, with those in the PP format being stronger than those on the online test, while the strongest one was found between the figure series completion and figure analogies tasks in both versions.

Table 7.3. Internal consistency indicated by Cronbach's alpha (α), McDonald's omega (ω) and the intercorrelations for the subtests.

	Online (N=337)					PP (N=378)				
	α	ω	FS	FA	NA	α	ω	FS	FA	NA
FS	.57	.62				.67	.75			
FA	.60	.76	.310			.57	.62	.503		
NA	.61	.69	.307	.293		.50	.55	.404	.370	
NS	.50	.53	.264	.322	.395	.69	.72	.435	.456	.446
All	.76	.80				.83	.85			

Note. p< .001 for all correlation coefficients.

Table 7.4 summarizes the psychometric properties of the IR test comparing the online and PP groups. In classical statistics, the percentage of correct answers and discrimination values are acceptable when they are higher than 0.3 (Ebel & Frisbie, 1991) and comparable between the two groups for most test items, except item 20. The results of the Rasch model analysis showed that the test items fitted the model to the data quite well in both versions. The infit for single items (weighted mean squares, MNSQ) ranged from 0.85 to 1.28 (Mean=0.99, SD=0.09) in the online group and from 0.87 to 1.20 (Mean=1.01, SD=0.11) in the paper-based group. The item difficulty ranged from −1.79 to 2.52 and from −1.57 to 2.16 on the online and the paper-based tests, respectively. Overall, these results suggest that the models were well supported with the empirical data in both test formats. Furthermore, the unidimensional model of the online sample resulted in a final deviance of 6,548.59 with 21 parameters (AIC=6,590.59, BIC=6,601.67), while that of the paper sample was estimated at a final deviance of 6,847.79 and 21 parameters (AIC=6,889.79, BIC=6,901.92). Consequently, the deviance, AIC and BIC of the online group were lower than those of the paper group, suggesting that the unidimensional model of the online sample fitted better to the empirical data than that of the PP group.

Table 7.4. Psychometric parameters of the IR test by modes of administration.

No.	Item	Correct answer		Discrimination		Difficulty		Infit	
		Online	PP	Online	PP	Online	PP	Online	PP
1	FS01	83.68	83.86	0.39	0.62	−0.84	−0.56	1.01	0.85
2	FS02	87.83	86.77	0.32	0.58	−1.23	−0.85	1.04	0.86
3	FS03	82.49	83.33	0.43	0.55	−0.74	−0.52	0.98	0.88
4	FS04	92.28	92.33	0.37	0.42	−1.79	−1.56	0.89	0.98
5	FS09	24.63	37.83	0.40	0.33	2.43	2.16	1.01	1.14
6	FA02	91.10	91.53	0.37	0.37	−1.62	−1.44	0.94	1.06
7	FA05	59.94	61.11	0.52	0.53	0.62	0.95	0.96	0.96
8	FA06	59.94	54.76	0.52	0.47	0.62	1.28	0.96	1.05
9	FA07	74.78	82.28	0.48	0.52	−0.20	−0.42	0.95	0.95
10	FA10	64.09	76.19	0.35	0.50	0.40	0.04	1.12	0.99
11	NA01	75.37	82.54	0.50	0.35	−0.24	−0.45	0.96	1.15
12	NA02	82.49	89.95	0.53	0.51	−0.75	−1.22	0.87	0.91
13	NA03	85.76	86.77	0.38	0.41	−1.03	−0.85	0.96	1.06
14	NA05	80.12	80.42	0.55	0.50	−0.56	−0.27	0.87	1.04
15	NA07	33.53	38.10	0.28	0.25	1.92	2.15	1.20	1.28
16	NS01	87.54	90.21	0.50	0.59	−1.20	−1.25	0.93	0.93
17	NS02	72.40	68.31	0.50	0.56	−0.06	0.48	0.96	1.00
18	NS03	60.53	70.11	0.45	0.50	0.59	0.44	1.02	1.07
19	NS05	48.66	59.52	0.42	0.56	1.17	1.03	1.05	0.98
20	NS07	23.15	62.96	0.22	0.53	2.52	0.85	1.18	0.98

7.4.3.2 A comparison online and PP versions by DIF analysis

DIF analysis was conducted with the online group set as a focal group. DIF analysis using Angoff's delta method with item purification parameters of the major axis (a=–0.4523, b=1.0925) in the last iterations and a detection threshold of 1.29 (significance level: 5%) suggested that 19 items displayed no DIF and only item 20 (NS07) was detected as a DIF item, these results thus favoring the paper group. Interestingly, most (7 out of 10) of the items with figure-related material yielded a positive effect size, while most (6 out of 10) of the items with number-related material had a negative effect size. In other words, the items comprising figures tended to favor the online group, while the items containing numbers seemed to favor the PP group, as depicted in Figure 7.1.

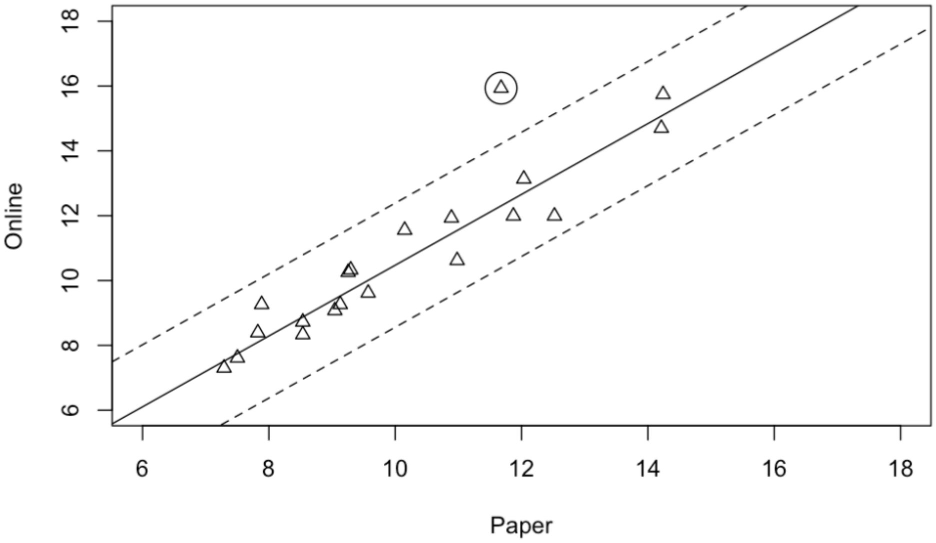


Figure 7.1. Delta plots for the dual modes of administration of the IR test.

Note: The items are located above the major axis, indicating that they are easier for the respondents in the reference group (PP group).

7.4.3.3 DBF analysis for comparison between two administration modes

For DBF analysis, the online group was also used as the focal group. Table 7.5 summarizes the results of the SIBTEST method at the task (item bundle) level. DBF analysis determined that significant DBF for the bundle of items was found in the figure analogies and figure series completion tasks, with the online group being favored. However, no DBF was indicated in the item bundle of the number series completion and number analogies tasks, although it seemed to favor the paper group. Generally, DBF analysis showed that the performance of the online group on the bundle of figure-related items was better than that of the PP group, whereas the results were apparently reversed in the number-related items.

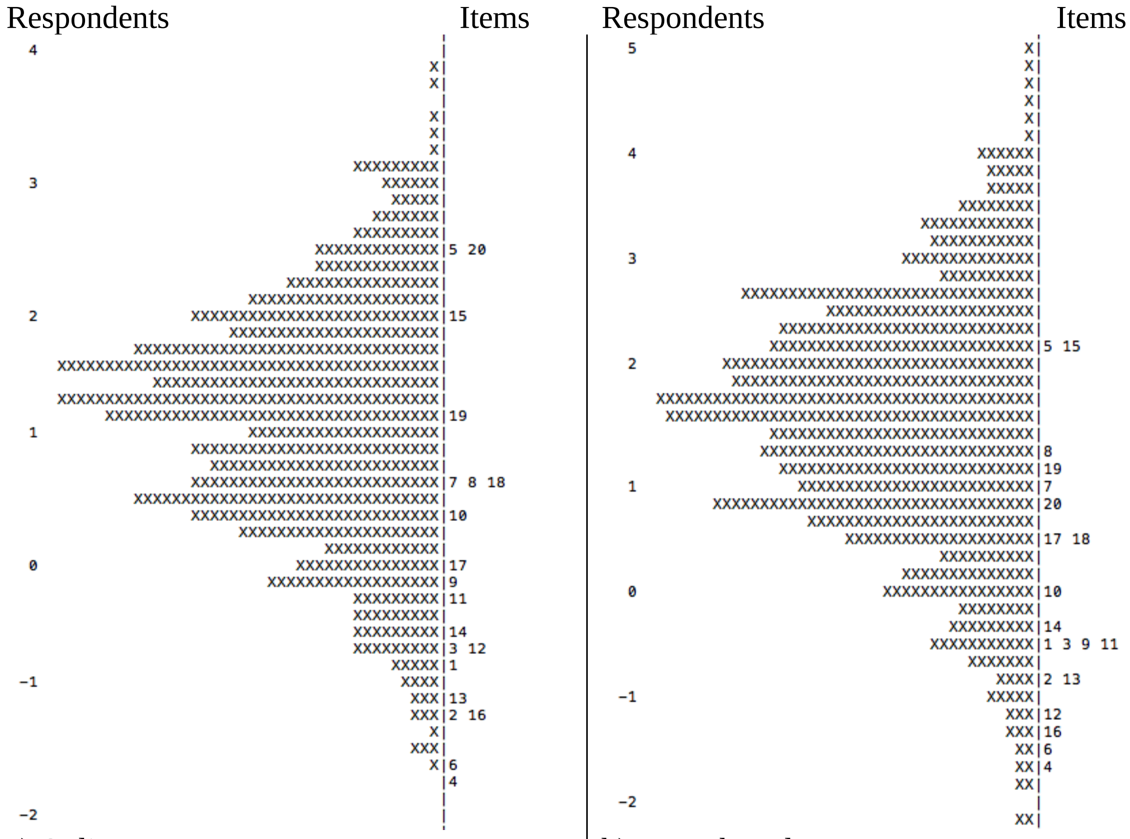
Table 7.5. Summary of the SIBTEST results at task level.

Item bundle	No. of items	β_S	p-value	Result favors
Figure analogies	5	−0.317	0.000	Online
Figure series completion	5	−0.455	0.000	Online
Number analogies	5	0.029	0.762	No DBF
Number series completion	5	0.095	0.078	No DBF

7.4.3.4 A descriptive comparison of the students’ performance

The average proficiency of the participants in the online group was 1.14 (SD=1.19) on the MLE scale of the Rasch measurement model, and that of the paper group was 1.56 (SD=1.39). In comparison with average item difficulty, the students' proficiency was estimated as higher than item difficulty (set at 0 logit). The students in the paper-based group found the test a bit easier than those in the online group. On average, the online participants completed 13.7 out of 20 items correctly (68.5%), while the students in the PP group did 14.8 out of 20 items correctly (74.4 %).

The Wright maps in Figure 7.2 present the patterns of the students' performance in the two groups. The results from the Wright map also suggest that the participants using the paper-based test medium tended to complete item 20 easily, as the results indicated the location of this item in the middle of the scale. However, this was not the case with the online participants because the map suggests that that item tended to be the most difficult one for them as it stood at the top of the scale. The finding is consistent with the result of Delta plots analysis, as discussed above. All items covered most of the participants' skills, but, in general, the test seemed easy for the students in both groups. All in all, the students' achievement on the online version was a bit lower than that of their peers on the PP version.



a) Online group   b) Paper-based group

Figure 7.2. The Wright maps for the online and the PP groups.

Note. Each "x" represents 0.6 cases.

Furthermore, we conducted a *t*-test to examine the difference in performance between the two groups. Table 7.6 provides a brief account of the students' performance on the test in the two modes of administration grouped by grade level. Generally, the average score of the paper-based group was higher than the mean score of the online group using Cohen's effect size value (d= 0.32), suggesting a small to medium significance. All the cohorts in the PP group scored higher than those in the online group, except in the 6th grade, where the students did the online test better than their peers who took the paper-based test.

Table 7.6. Students' performances on the IR test by modes of administration.

Grade	Online		PP		*t*	*p*	*Cohen's d*
	N	Mean (SD)	N	Mean (SD)			
6	44	1.07 (1.04)	59	−0.03 (1.01)	5.39	<.001	1.07
9	50	0.88 (1.23)	65	1.53 (1.23)	−2.78	.006	0.52
10	139	1.04 (1.18)	107	1.68 (1.27)	−4.02	<.001	0.52
11	104	1.41 (1.21)	147	2.12 (1.18)	−4.55	<.001	0.59
All	337	1.14 (1.19)	378	1.56 (1.39)	−4.33	<.001	0.32

7.5 Technology-based and paper-based testing in the CSVP test

7.5.1 Participants

The final data of this study was collected from 731 students from the 8th to 12th grades in nine secondary schools in Vietnam. As shown in Table 7.7, the mean ages of the students in the online and paper groups were 15.3 years and 15.5, respectively, and the male-to-female ratio seemed equivalent in each cohort and the whole sample. The students participated in the study voluntarily after the teachers introduced the aims of our research project. Most of the data were collected during March and April 2020. Because of the COVID-19 pandemic, some schooling activities were restricted, and thus the students took the test individually at home either in PP or online format without teacher supervision.

Table 7.7. Characteristics of the participants in the study 2.

Grade	N	Online/PP ratio (%)	Online		PP	
			Male/Female ratio (%)	Mean age (years)	Male/Female ratio (%)	Mean age (years)
8	150	52.0/48.0	54.7/45.3	13.1	50.0/50.0	13.6
9	144	50.7/49.3	50.7/49.3	14.6	50.7/49.3	14.6
10	235	48.9/51.1	47.3/52.7	15.8	50.0/50.0	15.8
11	129	31.0/69.0	33.3/66.7	16.7	29.3/70.7	16.8
12	73	58.9/41.1	55.2/44.8	17.7	61.4/38.6	17.8
Total	731	47.7/52.3	42.1/57.9	15.3	41.9/58.1	15.5

For the purposes of this study, we referred to physics grades in the previous semester to examine the equivalence of the prior abilities among the students in the two groups. Previous studies (Hejnová et al., 2018; Schwichow et al., 2020) have indicated that CVS and content knowledge in physics are closely tied, so we considered the results of the final test in physics in the previous semester as an index to support the assumption of prior ability equivalence between the two groups. Table 7.8 presents the summarized results of the *t*-test for comparing physics grades in the previous semester between the students in the online group and those in the PP group. Overall, there was no significant difference between the two groups in terms of physics achievement in the separate cohorts or in the entire sample.

Table 7.8. Comparison of students' achievement in physics in the previous semester.

Grade	Online		Paper		*t*	*p*
	N	Mean (SD)	N	Mean (SD)		
8	78	8.38(1.29)	72	8.13(1.35)	1.16	0.248
9	73	8.18(1.20)	71	7.86(1.18)	1.64	0.102
10	115	7.85(1.57)	120	7.60(1.69)	−0.17	0.247
11	40	7.93(1.53)	89	7.91(1.28)	0.07	0.947
12	40	7.73(1.72)	30	8.32(1.14)	−1.74	0.086
All	349	8.03(1.47)	382	7.88(1.42)	1.45	0.146

7.5.2 Instruments

The 24-item CVSP test was used to assess CVS in three sub-skills: *Identifying*, *Interpreting and Understanding*. The items were adapted from Schwichow et al. (2016), AAAS (2012), and TIMSS (1997) and other new items were developed by authors. The content of items related to the basic physics concepts (mechanics, heat and thermodynamics, and electricity and electromagnetism) in the secondary educational program in Vietnam. The items were formatted in the multiple-choice style, involving a stem with one correct answer and three distractors. We visualized graphical representations to minimize the influence of students' varying reading ability levels (see Section 6.3.1)

7.5.3 Findings of the study 2

7.5.3.1 Reliability and validity

The Cronbach' alpha (α), McDonald's omega (ω) and Pearson correlation among subskill tasks were shown in Table 7.9. Both two tests took an acceptable level in terms of internal consistency reliability, although online testing performed somewhat better in comparison to PP one. The strongest correlation was found between the identifying and interpreting tasks, followed by that between the interpreting and understanding tasks on both versions of the test.

Table 7.9. Internal consistency indicated by Cronbach's alpha (α), McDonald's omega (ω) and the intercorrelations for the subscales.

Sub-skill	Online (N=337)				PP (N=378)			
	α	ω	Identifying	Interpreting	α	ω	Identifying	Interpreting
Identifying	.64	.73			.67	.71		
Interpreting	.69	.75	.608***		.61	.68	.602***	
Understanding	.52	.63	.472***	.555***	.48	.58	.477***	.516***
All	.81	.84			.80	.82		

Note. *** p< .001.

Table 7.10 summarizes the psychometric properties of the CVSP test in the online and PP formats. As suggested by Ebel and Frisbie (1991), the discrimination indices were comparable for both test versions, but six items in the online version and four items in the paper one were still less than 0.3.

Table 7.10. Psychometric characteristics of the CVSP test by modes of administration.

No.	Item	Correct answer		Discrimination		Difficulty		Infit	
		online	PP	online	PP	online	PP	Online	PP
1	ID01	63.39	56.01	0.35	0.50	−0.97	−0.88	1.03	0.94
2	ID03	40.71	27.44	0.27	0.30	0.10	0.61	1.13	1.08
3	IN02	31.42	25.17	0.23	0.19	0.56	0.71	1.20	1.22
4	ID02	55.74	56.24	0.52	0.54	−0.61	−0.99	0.96	0.98
5	ID04	72.40	65.76	0.35	0.43	−1.48	−1.46	1.08	1.00
6	UN08	39.07	43.08	0.19	0.36	0.21	−0.32	1.22	1.12
7	IN05	39.62	25.62	0.36	0.42	0.20	0.71	1.08	0.99
8	IN07	39.07	37.64	0.58	0.44	0.20	−0.03	0.87	1.04
9	ID05	72.40	59.64	0.53	0.52	−1.50	−1.16	0.90	0.91
10	UN02	57.92	44.67	0.53	0.54	−0.71	−0.39	0.94	0.91
11	ID06	65.30	49.89	0.55	0.66	−1.06	−0.60	0.89	0.79
12	IN04	36.34	30.61	0.55	0.54	0.37	0.32	0.94	0.89
13	UN03	7.10	7.94	0.14	0.20	2.60	2.26	1.03	0.99
14	ID07	45.36	46.94	0.42	0.40	−0.12	−0.53	1.03	1.05
15	IN01	40.16	44.22	0.47	0.44	0.15	−0.44	1.01	1.05
16	IN03	55.46	43.99	0.60	0.59	−0.62	−0.29	0.90	0.86
17	UN06	21.58	19.50	0.52	0.58	1.20	1.03	0.85	0.83
18	IN08	68.31	54.20	0.52	0.54	−1.27	−0.82	0.92	0.91
19	UN07	3.55	7.48	0.27	0.14	3.47	2.34	0.98	0.96
20	UN01	51.37	43.31	0.58	0.32	−0.38	−0.21	0.89	1.13
21	UN04	29.78	20.18	0.39	0.39	0.67	1.03	1.02	0.98
22	ID08	52.46	41.04	0.52	0.51	−0.48	−0.29	0.92	0.94
23	UN05	66.94	64.40	0.11	0.02	−1.16	−1.28	1.24	1.35
24	IN06	30.33	24.72	0.57	0.39	0.62	0.68	0.85	1.03

Rasch model analysis also suggested that the model fitted the data well at item level for both test versions. The infit for single items ranged from 0.85 to 1.24 (Mean=1.00, SD=0.12) and from 0.79 to 1.35 (Mean=1.00, SD=0.12) for the online and the PP version, respectively. The item difficulty ranged from −1.50 to 3.47 in the online format and from −1.46 to 2.34 in the paper-based format. Moreover, the final deviance in the unidimensional model of the online sample is 9,479.76 with 25 parameters estimated (AIC=9,529.76, BIC=9,543.33), and this value in the paper sample is 10,372.01 with 25 parameters estimated (AIC=10,422.01, BIC=10,436.56). It is seen clearly that the deviance, AIC and BIC in the paper group were greater than those in the online group, suggesting that the unidimensional model fitted better to the online data than the PP one.

7.5.3.2 Measurement invariance with DIF analysis

Figure 7.3 depicts the results of DIF analysis applying Angoff's delta method when the online group was defaulted as a focal group. Examination of DIF operated with item purification parameters of the major axis: a= −3.04, b= 1.17, with a detection threshold of 1.03 and a significance level of 5%. The findings showed that there was no DIF item that was detected as a DIF item. Nevertheless, 10 items had a negative effect size, implying the results favor the PP group, while 14 items were estimated with a negative effect size, suggesting that suggesting that they were a bit easier for the online group.

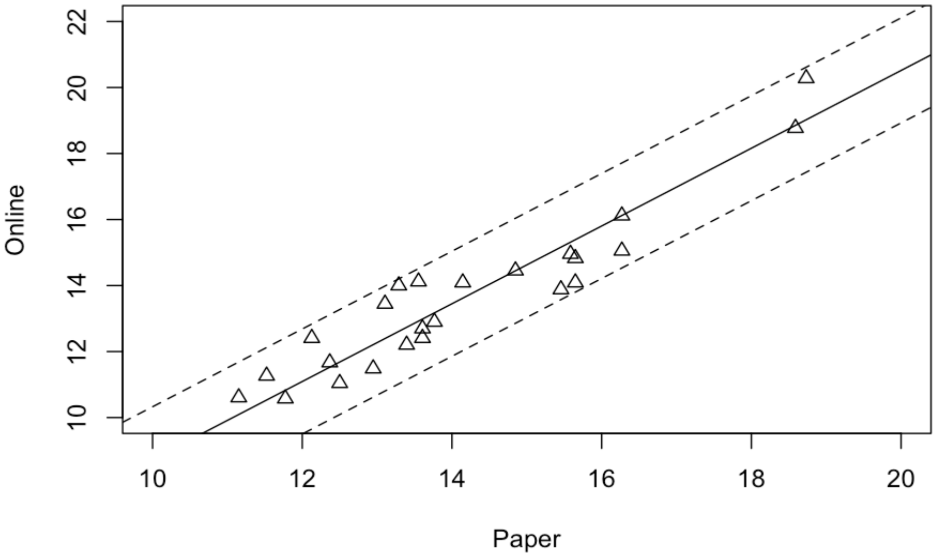


Figure 7.3. Delta plots for the dual administration modes of the CVSP test.

7.5.3.3 A comparison by DBF analysis in the CVSP test

The results of the DBF analysis were plotted in Table 7.11 when the online group was considered as the focal group with the SIBTEST method. Surprisingly, a significant DBF was indicated in all the bundles of CVS sub-skill items, with the results favoring the online group. In other words, the students who took the online test appeared to perform better than their peers who used the PP format in all the item bundles of the *identifying*, *interpreting* and *understanding* sub-skills.

Table 7.11. Summary of the SIBTEST results by subskills

Item bundle	No. of items	β_s	p-value	Result favors
Identifying	8	−0.709	0.000	Online
Interpreting	8	−0.654	0.000	Online
Understanding	8	−0.248	0.006	Online

7.5.3.4 Comparison of students' performance in dual media delivery modes

The online participants had an average proficiency of –0.33 digits (SD=1.13) and that of the PP group was –0.54 (SD=1.16). In comparison to item difficulty (Mean=0.00, SD=1.00), students' proficiency was estimated lower, suggesting that students felt the SVP test was somewhat difficult and that the paper group found the test more difficult than the online group did. On average, the online participants completed 10.89 out of 20 items (45.38%) correctly, while the paper group managed 9.63 items (40.13 %).

Additionally, the Wright maps in Figure 7.4 illustrate the contribution of students' performance in the two groups in relation to item difficulty. Though a general overview did not show significant differences between the two patterns on the maps, more online participants scored higher than 0 (digits). Items 13 and 19 were supposed to be the most difficult ones in both versions. Specifically, item 19 fell out of the spectrum of online test-takers. The locations of the other items covered the respondents' proficiency well, but the test seemed somewhat difficult for both student groups in general. Overall, students' achievement in the online version was higher than that for students who took the PP test.

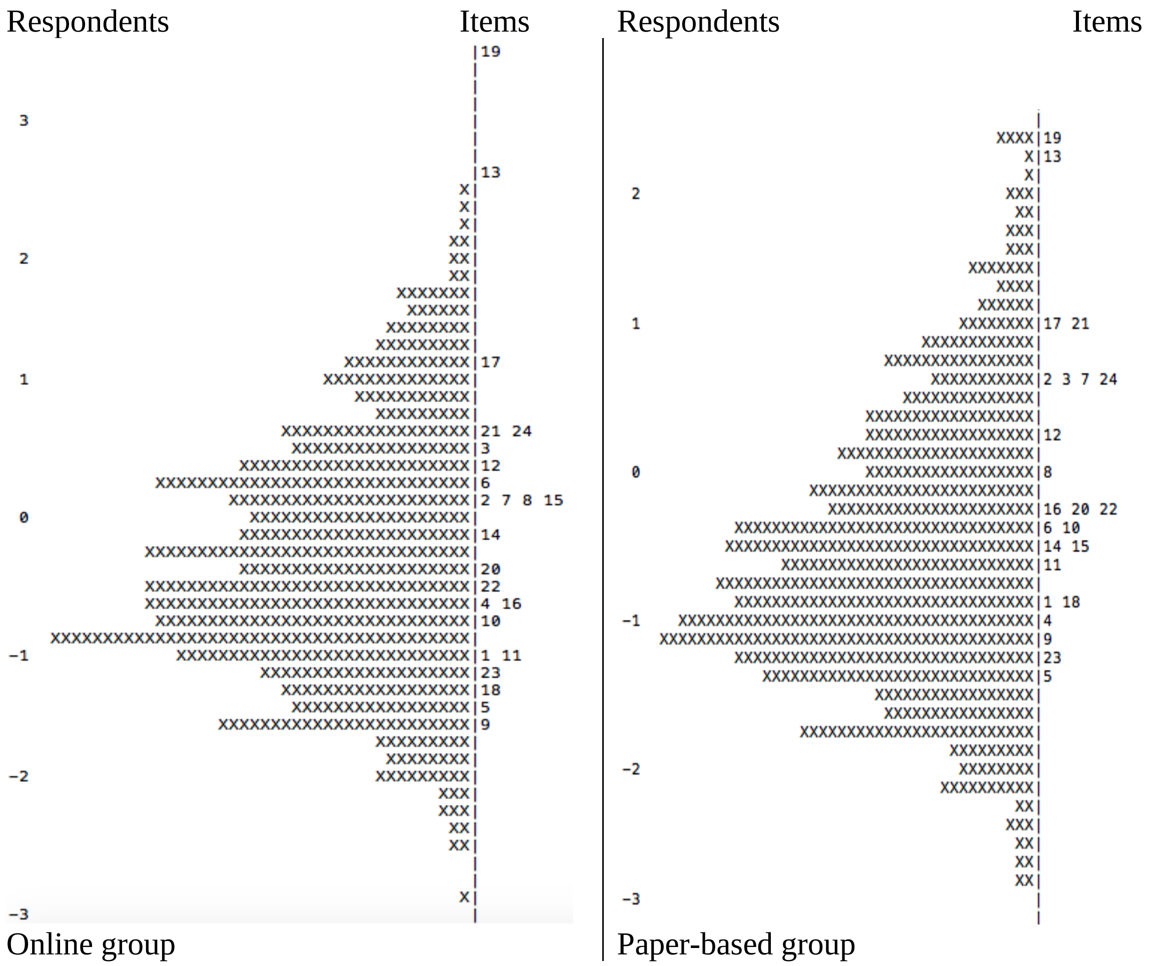


Figure 7.4. The Wright maps for the online and the paper-based groups on the CVS test. Note. Each "x" represents 0.6 cases.

Furthermore, the *t*-test was performed to compare students' performance between the two groups. Table 7.12 summarizes the results of the *t*-test on the CVSP test with regard to the two delivery modalities across the grade levels. Overall, the online group scored higher than the PP group with a small effect size with Cohen's value of 0.20. All the younger cohorts (the 8th, 9th and 10th grades) obtained significantly higher scores on the online version than their peers who took the traditional paper-based test. However, no significant disparity was found between the two 11th-grade groups, and even the PP group performed significantly higher than the online group in the 12th grade.

Table 7.12. Comparison of the student's performances on the CVSP test by delivery modality.

Grade	Online		PP		*t*	*p*	*Cohen's d*
	N	Mean (SD)	N	Mean (SD)			
8	78	−0.71(0.83)	72	−1.34(0.71)	4.96	<.001	0.81
9	73	−0.69(1.13)	71	−1.14(0.61)	3.01	.003	0.50
10	115	−0.15(1.07)	120	−0.74(0.82)	4.71	<.001	0.62
11	40	0.19(1.25)	89	0.26(1.02)	0.07	.758	0.03
12	40	0.21(1.21)	30	1.20(1.37)	−3.18	.002	0.77
All	349	−0.30(1.13)	382	−0.54(1.16)	2.80	.005	0.20

7.5.3.5 Additional analyses

The current version of the eDia platform allows us to measure the response time for each task as well as for a whole test. The average time to complete the test was 26.4 mins (SD= 10.5 mins). Figure 7.5 depicts the mean response time (seconds) in which online participants completed each item. The response time for most of the items (19 out of 24) ranged from 30 to 60 seconds, and two items required more than 60 seconds, while there were three items which students just took under 30 seconds each to handle. It seems that the average response time for an item in the *understanding* task was lower than other ones (see box plot in Figure 7.5). However, the result of the analysis of variance showed that response time did not significantly affect the execution of sub-skill tasks (F(2, 21)= 2.98, p= .073).

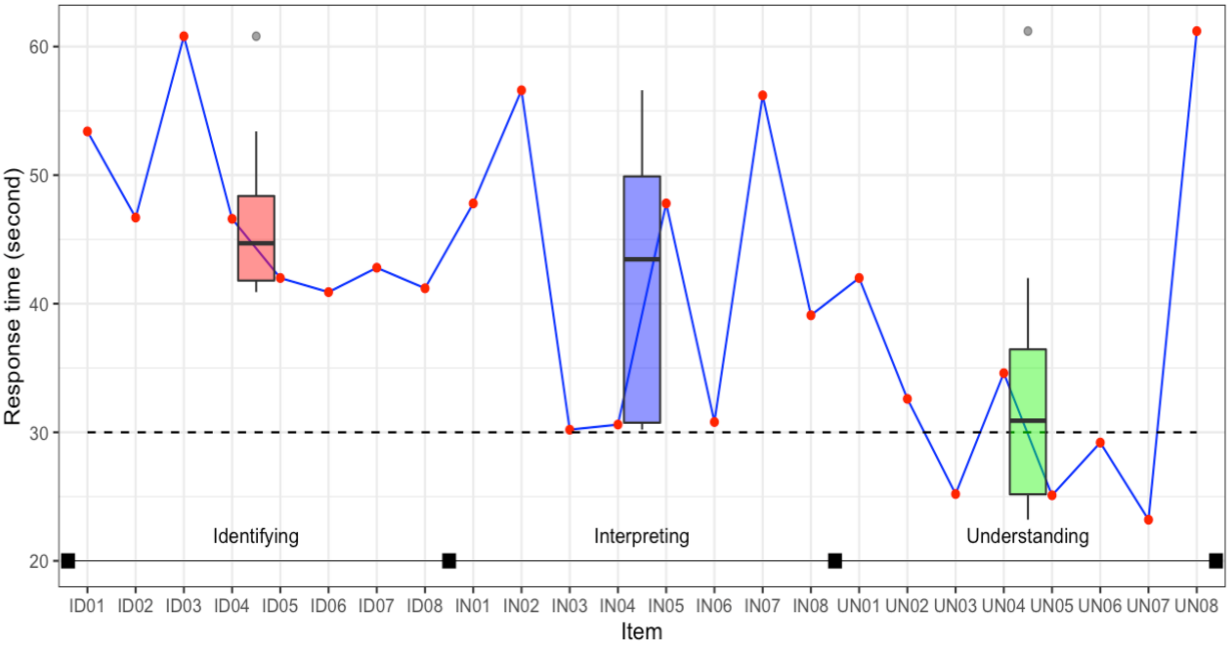


Figure 7.5. Response time for single items and boxplots for item bundles by subskills.

7.6 Discussion and conclusions

When tests evolve from paper-based assessment to TBA, equivalence is often expected no matter what the modes of administration are. Our study attempted to implement many-sided analyses to compare the performance of students across different grade levels, taking both traditional PP and online formats of the inductive reasoning test at the item, task and test levels into consideration. As regards the average total score, the results seemed to be more supportive of the PP version under teachers' supervision because the students performed better than their peers did via online assessment on the same test. These results were corroborated by the findings of previous studies (e.g., Schroeders & Wilhelm, 2010; Williamson et al., 2017), which concluded that the average performance of students on the inductive reasoning test was lower with digital forms of testing (e.g., using notebooks, tablets and smartphones) than with a traditional paper-based format. On the other hand, without supervision, the results seemed to favor the online format as the students who did the online testing performed better than their friends who participated in the PP version.

Interestingly, across delivery conditions, the results demonstrated that either with or without supervision, measurement with the Rasch model indicated that the online assessment had a better fit to the empirical data than the paper one. The results of the measurement invariance in DIF analysis with Angoff's delta approach showed the internal validity of the tests, thus demonstrating that they are acceptably comparable across the two modes of administration. The reliability and validity of the two tests were equivalent regardless of media delivery modalities (Hypothesis 4.1 is confirmed). These findings are in line with those of previous studies (e.g., Csapó et al., 2014; Hassler Hallstedt & Ghaderi, 2018; Schroeders & Wilhelm, 2010). However, item 20 (NS07) was identified as a DIF item with regard to the modes of administration in the first study. Possible reasons include the location of the item or students' fatigue or boredom. Item 20 is the last item ordered in the IR test and one of the most difficult items in the number series completion subtest, which might in turn make students lose interest in attempting the last one within a series of numbers on the screen. As a result, they might guess the answer rather than completing it with their own abilities and efforts. In addition, difficulty-based item order did not impact the item statistics on a paper-and-pencil multiple-choice test (e.g., difficulty, discrimination and point biserial) (Şad, 2020), but the item order may influence item properties on an online test. This issue raises an exciting issue for further research.

Furthermore, by employing the dimensionality-based approach to identifying clusters of items, we found that DBF favored the online group in the bundle of figure-related items. The findings did readly support the hypothesis H4.2. DBF analysis provided an insight into the tasks when constructing the tests. In the case of these cognitive tests, the study showed that students appeared to feel it was easier to complete tasks using pictures or graphs than those with simply numerical elements and words on the online test. This provided clear evidence of a link between test modality and the materials that made up the test items. It further illuminates how TBA can be more beneficial when items are composed of visually rich materials.

As regards the age groups of the respondents, the TBA format seemed to favor the younger cohorts over the older groups, while students at the upper grade levels who took the PP test clearly achieved higher scores than their peers who participated in the online tests. This may derive from the levels of digital competence in the younger and older generations, with the former enjoying the benefit of stronger digital skills than the latter (Oblinger & Oblinger, 2005). Such competence can affect students' performance on online format tests, since they are more familiar with the online format than their seniors.

Response time is an additional advantage of technology-based assessment. Investigation of response time enables us to clarify the solution behavior of test-takers that might effect bias in test performance (Wise & Kuhfeld, 2021). In this study, it seems that the items (UN03, UN05 and UN07) with an average response time of under 30 seconds have a low discrimination of less than 0.3 each. However, more studies should be conducted to examine the effects of time response on the psychometric properties of the items.

Although the investigations led to inconsistent findings regarding students' scores on the tests (The hypothesis H4.3 is not confirmed), the results showed the potential advantages of a technology-rich assessment when considering the psychometric characteristics of the tests, especially students' self-assessment process at home. Across the studies, the psychometric properties of the tests and individual items were comparable in dual administration modes and even leaning toward online versions. The study also provided multifaceted approaches to examining the testing equivalence and evidence for the feasibility of transferring from paper-and-pencil to TBA in the Vietnamese context. Future research needs to consider the possible factors relevant to test modality, familiarity with electronic devices (Schroeders & Wilhelm, 2010), and administration procedures. More sophisticated comparative studies should continue to explore this as advanced electronic devices are becoming increasingly essential tools both at school and in the home. The issue of test fairness will continue to be considered as an important construct in the upcoming decade (Iliescu & Greiff, 2019) and the COVID-19 pandemic is currently a significant challenge for traditional modes of delivery.

CHAPTER 8. GENERAL DISCUSSION AND CONCLUSIONS

8.1 Discussion and Conclusions

Based on the PRISMA guidelines, the systematic review established a bridge between theoretical concepts and empirical paradigms with a particular emphasis on the construction of the instruments, psychometric issues and technology approaches to assess reasoning capacities in school settings. Among several forms of reasoning, IR and SR are increasingly interested in educational contexts. IR, a central component of fluid intelligence (Perret, 2015) and SR, a foundational pillar of scientific literacy (Lawson, 2009), have been considered as core competence to succeed in 21st-century life. We found that the types of reasoning tasks appeared unchanged over twenty-three years, but they grew more diverse and gradually evolved from PP to online formats available in smartphones, tablets, and wearable devices. Using a touch screen or drag-and-drop method is a potential approach for developing the test item. Such new technology allows us to create dynamic problem tasks, where children can easily respond to interactive tasks by touching a screen during the adaptive tests, dynamically functioning based on the Bayesian network technique (Mahnane & Hafidi, 2016). Such tasks are not only designed to develop children’s thinking, but also to enhance their mathematics and digital competencies. The literature review also revealed that reasoning skills are closely associated with several variables, such as age, discipline performance, parental factors and problem-solving skills, while gender differences may depend on particular cultures. The findings provided an orientated vision for the assessment of reasoning skills in future and inspired us to project three empirical studies in Vietnamese context.

The empirical studies confirmed the quality of the adapted tests and questionnaires by undergoing the investigations of validity and reliability across grade levels. We also developed a new test to measure scientific reasoning in CVS in physics for secondary school students in Vietnamese context. The initial analysis showed that the test served as a reliable instrument tool to assess the CVS capacity in Vietnamese students. Equivalences are often expected between different delivery modalities and gender; and thus, our studies also used multifaceted methods to observe measurement invariance of the tests and questionnaires. The results demonstrated that no DIF was found in the test instruments with respect to administration modes and gender. Nevertheless, the DBF analyses resulted in supporting the online format on the item bundles constructed of figure-related resources. Online format appears to be more beneficial when items are composed of visually rich materials. Response time was also analysed as a plus of TBA accompanying the test development process. The feasibility of transferring from PP to TBA was examined through the empirical studies. The findings suggested that the results from the online formats showed a better fit to the data than those from PP ones. Particularly, within a direct supervision of teachers, the PP version proved to be better than the online version regarding students' scores, but the contrasting findings were found in the testing occasions without the direct supervision of teachers. Surprisingly, the online versions always had better psychometric properties, compared to the PP ones in these investigations. Therefore, the studies provided evidence of a possibility of applying TBA in Vietnam with various potential technical analyses which are in line with the development of educational instruments.

Students' performance on the reasoning tests increased grade by grade during secondary school education. The developmental curves of IR and SR in students showed similarly across the grade cohorts, in which the most rapid development was observed between the age group of 12 - 14 years (6th - 8th grades), but the growth rate apparently slowed down after those years. These results are mostly in line with the findings of Csapó (1997), Molnár and Csapó (2011), Molnár et al. (2013), Díaz-Morales and Escribano (2013) and Csapó et al. (2019). Meanwhile, the development of children's CVS fitted very well to the symmetric log-logistic curve, and the fastest growth was detected at the age range of the 14 - 16 years (the 10th grade). The findings are consistent with the results of previous studies (e.g. Han, 2013; Schwichow et al., 2020; Zimmerman, 2007). However, the changing patterns of SM in children apparently fluctuated and even reduced in some motivational components across grade cohorts. Specially, students' scores on self-efficacy and active learning strategy scales reduced drastically in the older groups. Regarding motivation in learning physics, the changing patterns of children's motivation showed the similar trend with the science motivation in general, but the reduction rate tended to be quite lower and levelled-off throughout grade levels. Likewise, among physics motivation scales, the findings found that children scored lowest points in the scale of the confidence in learning physics.

Concerning gender differences, although the mean scores of boys were slightly higher than those of girls on all the reasoning tests, no significant difference was found in any the reasoning test. These results were in agreement with former studies (e.g. Díaz-Morales & Escribano, 2013; Kambeyo, 2018; Mayer et al., 2014; Piraksa et al., 2014), but these were inconsistent with some other findings; for instance, girls attained significantly higher scores than boys in IR (Díaz-Morales & Escribano, 2013) or boys performed better than girls in SM tests (Tairab, 2015; Tekkaya & Yenilmez, 2006; Valanides, 1997). However, there was a statistically significant difference in understanding the indeterminacy of confounded experiments subskill of CVS in the current research, in which males performed better than females. For science motivation, girls reported higher scores than boys in most of the subscales in the SMTSL questionnaire, but a significant difference was only indicated in the achievement goals subscale (favouring for girls). In specific-domain subject, a gender difference was found in the students' attitude and motivation toward learning physics, in which boys were more motivated than girls in all three subscales.

Furthermore, the findings revealed that cognitive abilities, motivation, and parental factors are the important predictors of children's STEM achievement. The findings seemed to be in line with the existing studies (e.g. Fan & Williams, 2010; Steinmayr & Spinath, 2009) and meta-analysis review by Kriegbaum et al. (2018). Parents' education levels and parental involvement in schooling impacted meaningfully on their kids' performance in schools. Particularly, the father's education level was denoted as a highest predictor of IR and STEM achievement as well, while the mother's education level played as an indirect factor in predicting children's motivation and STEM performance through the parental involvement variable. The mother's and father's education levels were proved as educational resources at home, so these variables were strongly associated to science achievement (OECD, 2017b). Ample parental support and engagement as well as interest in school activities play a key role in children's motivation in learning and engaging in school activities, which has been confirmed in previous studies (e.g. Fan & Williams, 2010; Fan et al., 2012; Gonzalez-DeHass et al., 2005; OECD, 2017b). Likewise, parental involvement in schoolwork is a significant factor explaining the success of children in learning STEM subjects. These results have seemingly corresponded to those of the study by Ganzach (2000), and this may derive from the typical culture in Vietnam, where parents tend to consider their kids' performance in schools (Hoang et al., 2014; Phan, 2004).

The acquisition of knowledge in individual disciplines contributes importantly to the development of the domain-general processes in school-age children. The study showed a close association between two forms of reasoning skills and between CVS and the content knowledge test, suggesting that teaching reasoning through specific science content knowledge can contribute to developing the general thinking skills of the students. This was in line with the existing studies (e.g. Kambeyo, 2018; Korom et al., 2017; Schwichow et al., 2020), which found a strong, and positive correlation between IR and SR. General thinking can enrich specific-domain thinking skills and enhance children's success in learning content knowledge, and in reverse, children can develop their skills through learning individual subjects. These results also confirmed that thinking can be taught either explicitly in specific courses or embedded in the regular school curricula within the framework of school disciplines (Csapó, 1999), and the study partly provided evidence that students' reasoning capacities can be assessed, and their development can be monitored as the way it is practiced in several countries (Vainikainen, 2014). In addition, this possibility paves the way for estimating the impact of the teaching and learning activities of the current core curriculum on thinking skills and assessing the effect of any curricular change in the future.

8.2 Educational implication

The findings of the cross-sectional studies have contributed to an increased understanding of development of reasoning capacities in school-age children, the latent factors predicting these abilities and their roles in learning science subjects in schools.

The results revealed that general reasoning proficiencies in children develop most rapidly at the middle school level, suggesting that lower secondary school is the most appropriate time to nurture thinking skills in the school curricula. Thus, parents and teachers should be aware of the golden opportunity to boost children's reasoning skills in those years by teaching school subjects and practising daily activities which can enhance their reasoning skills. Because both physical and social experience influence cognitive ability, teachers can employ inquiry-based methods through science topics to strengthen both general cognitive skills and knowledge gains. Moreover, girls tended to develop their reasoning skills earlier than boys at secondary education level, but the growth seemed to level off during high school years. Teachers and parents should recognise this trend to find more appropriate facilities to support students' improvement during those years.

Generally, students were highly motivated to learn science in general and particularly to learn physics. Students scored high points for the active learning strategies subscale, but a remarkable reduction was observed at the older age groups in this scale. Students' scores also decreased significantly in the self-efficacy scale across grade levels. In other words, students' beliefs about their own capacity to apply the various strategies to acquire new knowledge appeared to drop gradually across grade levels. Notably, most of the students in all grade cohorts reported lowest scores in the learning environment stimulation scale. Similarly, students' confidence in learning physics were quite low indicated by the lower mean score, compared with other scales. Correspondingly, both schools and families should be concerned about how to inspire children's motivation to learn science. Teachers, school leaders and school psychologists should be aware of these trends to find more supportive facilities to improve learning environments and enhance children's confidence in learning science at schools. The findings are important for identifying the priority factors in enhancing science motivation in school practice in Vietnam.

Furthermore, our proposed test which was developed to assess scientific reasoning in CVS in physics for secondary school students not only provided an overall construction of problem tasks for measuring CVS, but also acknowledged the important role of CVS skill in learning physics and science. The findings of the investigations of the latent factors impacting on item difficulty in the CVSP test may provide helpful information for test developers and teachers who consider to assess students' CVS capacity.

TBA offers several solutions for measuring a broader competency range when using an adaptive testing and a fix-length format by using touch screen method (Molnár & Csapó, 2011). The findings showed that the psychometric properties of the test and student's achievements tended to be higher on the items with a rich-visualization representation than the items constructed of textual and numerical materials. This suggests that applying advanced technology in developing test items is more meaningful when we utilise the multimedia effectively for problem tasks with rich-visualised representation. Our project was the first major project which considered assessing a general cognitive skill with online instruments in Vietnamese context. The practicality of using dual media administration across these studies may encourage other studies as well as the use of computers for learning and testing purposes. Therefore, our project is expected to offer insights into the students' cognitive development and how thinking skills should be enhanced through the curriculum which the stakeholders including policymakers, teachers and parents may consider in future.

8.3 Limitations and future directions

Some limitations may be acknowledged in these investigations. The theoretical groundings are endeavoured based on the popular concepts of reasoning and motivation which may be a controversial topic in the psychological field. New concepts can be formed in new contexts; consequently, the definition and classification of reasoning capacity and relevant aspects presented in our review represent one of several approaches. Some issues can come from the measures. For example, the IR tests we used in our studies consist of four kinds of non-verbal tasks that may not cover the full scope of reasoning proficiency. Although non-verbal problem tasks minimise the effects of culture contexts and language translation issues, verbal problem tasks play a critical role in learning, but these studies cannot deal with such matters. Future research should consider the space when assessing students' reasoning capacity. It is the first time that the SR and CVSP tests were used to assess students in the Vietnamese context, so some items need to be revised in order to improve validity and reliability for next-generation versions. Moreover, the motivation questionnaires may not cover all the motivational factors. The next-generation instruments should be modified continually in the Vietnamese version with a larger sample size.

Other concerns may also arise from the study sample which was conducted in An Giang province (Vietnam) with cross-sectional investigations. Therefore, the generalization of the findings should be under caution. Future studies should consider recruiting larger samples representing different educational contexts. More large-scale assessment and longitudinal approach need to be managed to weigh the success of the current curricula.

The application of advanced technology is quite sophisticated and diverse; we have merely noted a few features without covering all aspects of the assessment of reasoning. For example, the CVSP test did not consider assessing the *Planning controlled experiments* skills of students. In principle, developing the kind of problems in the virtual environment is possible, using simulated real-life experimental tasks in advance (Luecht & Sireci, 2011). However, the successful adaptation of the test instruments may inspire other researchers to adapt others in future.

All of all, we hope that researchers, educators, teachers, and parents can find useful information about the changing patterns of reasoning skills and motivation and practical means to effectively promote them in school-year children from our project.

REFERENCES

Acharya, N., & Joshi, S. (2009). Influence of parents' education on achievement motivation of adolescents. *Indian Journal Social Science Researches*, *6*(1), 72–79.

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

Adams, R., & Wu, M. (2010). Modelling polytomously scored items with the rating scale and partial credit models. *Letzter Zugriff Am*, *30*, 2015.

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

Adey, P., Csapó, B., Demetriou, A., Hautamäki, J., & Shayer, M. (2007). Can we be intelligent about intelligence?. Why education needs the concept of plastic general ability. *Educational Research Review*, *2*(2), 75–97. https://doi.org/10.1016/j.edurev.2007.05.001

Adey, P., & Shayer, M. (1994). Really raising standards. In *Really Raising Standards: cognitive inrtervention and academic achievement*. Routledge. https://doi.org/https://doi.org/10.4324/9780203137284

Al-Balushi, S. M., Al-Musawi, A. S., Ambusaidi, A. K., & Al-Hajri, F. H. (2017). The effectiveness of interacting with scientific animations in Chemistry using mobile devices on grade 12 students' spatial ability and scientific reasoning skills. *Journal of Science Education and Technology*, *26*(1), 70–81. https://doi.org/10.1007/s10956-016-9652-2

American Association for the Advancement of Science (AAAS). (2012). *AAAS Science Assessment - Project2061*. https://www.aaas.org/programs/project-2061

American Education Research Association, American Psychological Association, and N. M. in E. (1999). *Standards for educational and psychological testing*. American Psychological Association. https://www.aera.net/Portals/38/1999 Standards_revised.pdf

Anderman, E. M., & Dawson, H. (2011). Learning with motivation. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 219–242). Taylor & Francis.

Andressa, H., Mavrikaki, E., & Dermitzaki, I. (2016). Adaptation of the Students' Motivation Towards Science Learning Questionnaire To Measure Greek Students' Motivation Towards Biology Learning. *International Journal Of Biology Education*, *4*(2). https://doi.org/10.20876/ijobed.56334

Andrich, D., Marais, I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social and health sciences*. Springer.

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berck (Ed.), *Handbook of methods for detecting item bias* (pp. 96–116). Baltimore, MD: Johns Hopkins University Press.

Ariës, R. J., Ghysels, J., Groot, W., & Brink, H. M. Van Den. (2016). Combined working memory capacity and reasoning strategy training improves reasoning skills in secondary social studies education: Evidence from an experimental study. *Thinking Skills and Creativity*, *22*(2016), 233–246. https://doi.org/http://dx.doi.org/10.1016/j.tsc.2016.10.008

Bailey, S. K. T., Neigel, A. R., Dhanani, L. Y., & Sims, V. K. (2018). Establishing Measurement Equivalence Across Computer- and Paper-Based Tests of Spatial Cognition. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *60*(3), 340–350. https://doi.org/10.1177/0018720817747731

Baltes, P. B. (1968). Longitudinal and cross-sectional sequences in the study of age and generation effects. *Human Development*, *11*(3), 145–171.

Bao, L., Fang, K., Cai, T., Wang, J., Yang, L., Cui, L., Han, J., Ding, L., & Luo, Y. (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

Bao, L., Xiao, Y., Koenig, K., & Han, J. (2018). Validity evaluation of the Lawson classroom test of scientific reasoning. *Physical Review: Physics Education Research*, *14*(2), 1–19. https://doi.org/10.1103/PhysRevPhysEducRes.14.020106

Barkl, S., Porter, A., & Ginns, P. (2012). Cognitive training for children: effects on inductive reasoning, deductive reasoning, and mathematics achievement in an australian school setting. *Psychology in the Schools*, *49*(9), 828–842. https://doi.org/10.1002/pits

Bathgate, M., & Schunn, C. (2017). The psychological characteristics of experiences that influence science motivation and content knowledge. *International Journal of Science Education*, *39*(17), 2402–2432. https://doi.org/10.1080/09500693.2017.1386807

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

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-ricci, M., & Rumble, M. (2012). Defining Twenty-First Century Skills. In *Assessment and Teaching of 21st*

Century Skills. Springer Science+Business Media. https://doi.org/10.1007/978-94-007-2324-5

Blum, D., Holling, H., Galibert, M. S., & Forthmann, B. (2016). Task difficulty prediction of figural analogies. *Intelligence, 56*, 72–81. https://doi.org/10.1016/j.intell.2016.03.001

Boğar, Y. (2019). Evaluation of the scientific reasoning skills of 7th grade students in science course. *Universal Journal of Educational Research*, *7*(6), 1430–1441. https://doi.org/10.13189/ujer.2019.070610

Bonney, C. R., & Sternberg, R. J. (2011). Learning to think critically. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of Research on learning and Instruction*. Taylor & Francis.

Boroş, D., & Sas, C. (2011). Developing reasoning in students with above average cognitive skills. *Journal of Psychological and Educational Research, 19*(3), 54–66.

Bouffard, T., Boileau, L., & Vezeau, C. (2001). Students' transition from elementary to high school and changes of the relationship between motivation and academic performance. *European Journal of Psychology of Education*, *16*(4), 589–604. https://doi.org/10.1007/BF03173199

Boujaoude, S., Salloum, S., & Abd-El-Khalick, F. (2007). Relationships between selective cognitive variables and students' ability to solve chemistry problems. *International Journal of Science Education*, *26*(1), 63–84. https://doi.org/10.1080/0950069032000070315

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

Bühner, M., Kröner, S., & Ziegler, M. (2008). Working memory, visual-spatial-intelligence and their relationship to problem-solving. *Intelligence, 36*(6), 672–680. https://doi.org/10.1016/j.intell.2008.03.008

Buerger, S., Kroehne, U., Koehler, C., & Goldhammer, F. (2019). What makes the difference? The impact of item properties on mode effects in reading assessments. *Studies in Educational Evaluation, 62*(April), 1–9. https://doi.org/10.1016/j.stueduc.2019.04.005

Burger, K. (2010). How does early childhood care and education affect cognitive development? An international review of the effects of early interventions for children from different social backgrounds. *Early Childhood Research Quarterly, 25*(2), 140–165. https://doi.org/10.1016/j.ecresq.2009.11.001

Carroll, J. B. (1993). *Human cognitive abilities: A survay of factor-analytic studies*.

Cambridge University Press.

Cavas, P. (2011). Factors affecting the motivation of Turkish primary students for science learning. *Science Education International*, *22*(1), 31–42.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6). https://doi.org/10.18637/jss.v048.i06

Chan, Y. L., & Norlizah, C. . (2018). Students' motivation towards science learning and students' science achievement. *International Journal of Academic Research in Progressive Education and Development*, *6*(4), 174–189. https://doi.org/10.6007/IJARPED/v6-i4/3716

Chapman, J. W. (1988). Cognitive-motivational characteristics and academic achievement of learning disabled children: A longitudinal study. *Journal of Educational Psychology*, *80*(3), 357–365. https://doi.org/10.1037/0022-0663.80.3.357

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

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif : An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/Item response theory and monte carlo simulations. *Journal of Statistical Software*, *39*(8). https://doi.org/10.18637/jss.v039.i08

Chraif, M., & Dumitru, D. (2015). Differences between motivation from competition and motivation from individual goals under the influence of inductive reasoning. *Procedia - Social and Behavioral Sciences*, *187*(2015), 745–751. https://doi.org/10.1016/j.sbspro.2015.03.157

Chu, S. K. W., Reynolds, R. B., Tavares, N. J., Notari, M., & Lee, C. W. Y. (2017). 21st century skills development through inquiry-based learning. In *21st Century Skills Development Through Inquiry-Based Learning: From Theory to Practice* (Issue January). Springer Singapore. https://doi.org/10.1007/978-981-10-2481-8

Chuang, M. H., & She, H. C. (2013). Fostering 5 th grade students' understanding of science via salience analogical reasoning in on-line and classroom learning environments. *Educational Technology and Society*, *16*(3), 102–118.

Clark, M. H., Middleton, S. C., Nguyen, D., & Zwick, L. K. (2014). Mediating relationships between academic motivation, academic integration and academic performance. *Learning and Individual Differences*, *33*, 30–38.

https://doi.org/10.1016/j.lindif.2014.04.007

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

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

Csapó, B. (1999). Improving thinking through the content of teaching. In J. H. & & B. C. M. Hamers, J. E. H. van Luit (Eds.), *Teaching and learning thinking skills* (pp. 37–63). Swets & Zeitlinger.

Csapó, B. (2001). Az induktív gondolkodás fejlődésének elemzése országos reprezentatív felmérés alapján [An analysis of the development of inductive reasoning on the basis of a large-scale survey]. *Magyar Pedagógia*, *101*(3), 373–391.

Csapó, B., Ainley, J., Bennett, R. E., Latour, T., & Law, N. (2012). Technological issues for computer-based assessment. In *Assessment and Teaching of 21st Century Skills* (pp. 143–230). Springer Netherlands. https://doi.org/10.1007/978-94-007-2324-5_4

Csapó, B., Hotulainen, R., Pásztor, A., & Molnár, G. (2019). Az induktív gondolkodás fejlődésének összehasonlító vizsgálata: online felmérések Magyarországon és Finnországban [A comparative study of the development of inductive thinking: online surveys in Hungary and Finland]. *Neveléstudomány [Educational Science: Education Research Innovation]*, *7*(3–4), 5–24.

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., Molnár, G., & Nagy, J. (2014). Computer-based assessment of school readiness and early reasoning. *Journal of Educational Psychology*, *106*(3), 639–650. https://doi.org/10.1037/a0035756

Datu, J. A. D., King, R. B., & Valdez, J. P. M. (2018). Psychological capital bolsters motivation, engagement, and achievement: Cross-sectional and longitudinal studies. *The Journal of Positive Psychology*, *13*(3), 260–270. https://doi.org/10.1080/17439760.2016.1257056

de Koning, E., Hamers, J. H. M., Sijtsma, K., & Vermeer, A. (2002). Teaching inductive reasoning in primary education. *Developmental Review*, *22*(2), 211–241. https://doi.org/10.1006/drev.2002.0548

Demetriou, A., Spanoudis, G., & Mouyi, A. (2011). Educating the developing mind: Towards an overarching paradigm. *Education Psychology Review*, *23*(2011), 601–663. https://doi.org/10.1007/s10648-011-9178-3

Dermitzaki, I., Stavroussi, P., Vavougios, D., & Kotsis, K. T. (2013). Adaptation of the students' motivation towards science learning (SMTSL) questionnaire in the Greek language. *European Journal of Psychology of Education*, *28*(3), 747–766. https://doi.org/10.1007/s10212-012-0138-1

Díaz-Morales, J. F., & Escribano, C. (2013). Predicting school achievement: The role of inductive reasoning, sleep length and morningness-eveningness. *Personality and Individual Differences*, *55*(2), 106–111. https://doi.org/10.1016/j.paid.2013.02.011

DiCerbo, K. E., Xu, Y., Levy, R., Lai, E., & Holland, L. (2017). Modeling student cognition in digital and nondigital assessment environments. *Educational Assessment*, *22*(4), 275–297. https://doi.org/10.1080/10627197.2017.1382343

Dilivio, L. L. (2009). *Relationships between inductive reasoning and knowledge-based variables: an integration of psychometric and cognitive approaches* (Issue file:///Users/arieljames/Downloads/1170588.pdf). The University at Buffalo, the State University of New York.

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

Ding, L., Wei, X., & Mollohan, K. (2016). Does higher education improve student scientific reasoning skills? *International Journal of Science and Mathematics Education*, *14*(4), 619–634. https://doi.org/10.1007/s10763-014-9597-y

Dorfman, B. S., & Fortus, D. (2019). Students' self-efficacy for science in different school systems. *Journal of Research in Science Teaching*, *January*, 1037–1059. https://doi.org/10.1002/tea.21542

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement*, *33*(4), 465–484. https://doi.org/10.1111/j.1745-3984.1996.tb00502.x

Du, N. N. (2015). *Factors influencing teaching for critical thinking in Vietnamese lower secondary schools: a mixed method study focussed on history (Unpublished doctoral dissertation)*. Newcastle University.

Duckworth, A. L., Grant, H., Loew, B., Oettingen, G., & Gollwitzer, P. M. (2011). Self-regulation strategies improve self-discipline in adolescents: benefits of mental contrasting and implementation intentions. *Educational Psychology, 31*(1), 17–26. https://doi.org/10.1080/01443410.2010.506003

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (Vol. 11, Issue 2). Prentice-Hall.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Edelsbrunner, P. A. (2017). *Domain-general and domain-specific scientific thinking in childhood: Measurement and educational interplay (Unpublished doctoral dissertation)*. https://doi.org/https://doi.org/10.3929/ethz-b-000247401

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

Engelmann, K., Neuhaus, B. J., & Fischer, F. (2016). Fostering scientific reasoning in education–meta-analytic evidence from intervention studies. *Educational Research and Evaluation*, *22*(5–6), 333–349. https://doi.org/10.1080/13803611.2016.1240089

Evans, J. S. B. T. (1993). The cognitive psychology of reasoning: An introduction. *The Quarterly Journal of Experimental Psychology Section A*, *46*(4), 561–567. https://doi.org/10.1080/14640749308401027

Fan, W., & Williams, C. M. (2010). The effects of parental involvement on students' academic self-efficacy, engagement and intrinsic motivation. *Educational Psychology, 30*(1), 53–74. https://doi.org/10.1080/01443410903353302

Fan, W., Williams, C. M., & Wolters, C. A. (2012). Parental involvement in predicting school motivation: Similar and differential effects across ethnic groups. *The Journal of Educational Research*, *105*(1), 21–35. https://doi.org/10.1080/00220671.2010.515625

Fischer, A., Greiff, S., & Funke, J. (2012). The process of solving complex problems. *The Journal of Problem Solving*, *4*(1). https://doi.org/10.7771/1932-6246.1118

Gagné, F., & St Père, F. (2001). When IQ is controlled, does motivation still predict achievement? *Intelligence, 30*(1), 71–100. https://doi.org/10.1016/S0160-

2896(01)00068-X

Ganzach, Y. (2000). Parents' education, cognitive ability, educational expectations and educational attainment: Interactive effects. *British Journal of Educational Psychology*, *70*(3), 419–441. https://doi.org/10.1348/000709900158218

Garcia, T., & Pintrich, P. R. (1995). The role of possible selves in adolescents' perceived competence and self-regulation. *Annual Meeting of the American Research Association*.

Gates, N. J., & Kochan, N. A. (2015). Computerized and on-line neuropsychological testing for late-life cognition and neurocognitive disorders. *Current Opinion in Psychiatry*, *28*(2), 165–172. https://doi.org/10.1097/YCO.0000000000000141

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

Gerber, B. L., Cavallo, A. M. L., & Marek, E. A. (2001). Relationships among informal learning environments, teaching procedures and scientific reasoning ability. *International Journal of Science Education*, *23*(5), 535–549. https://doi.org/10.1080/09500690116971

Glynn, S. M., Brickman, P., Armstrong, N., & Taasoobshirazi, G. (2011). Science motivation questionnaire II: Validation with science majors and nonscience majors. *Journal of Research in Science Teaching*, *48*(10), 1159–1176. https://doi.org/10.1002/tea.20442

Glynn, S. M., Taasoobshirazi, G., & Brickman, P. (2009). Science motivation questionnaire: Construct validation with nonscience majors. *Journal of Research in Science Teaching*, *46*(2), 127–146. https://doi.org/10.1002/tea.20267

Gonida, E. N., & Urdan, T. (2007). Parental influences on student motivation, affect and academic behaviour: Introduction to the Special Issue. *European Journal of Psychology of Education*, *22*(1), 3–6. https://doi.org/10.1007/BF03173685

Gonzalez-DeHass, A. R., Willems, P. P., & Holbein, M. F. D. (2005). Examining the relationship between parental involvement and student motivation. *Educational Psychology Review*, *17*(2), 99–123. https://doi.org/10.1007/s10648-005-3949-7

Gottfried, A. E., Fleming, J. S., & Gottfried, A. W. (2001). Continuity of academic intrinsic motivation from childhood through late adolescence: A longitudinal study. *Journal of Educational Psychology*, *93*(1), 3–13. https://doi.org/10.1037/0022-0663.93.1.3

Gough, D. (2007). Weight of evidence: A framework for the appraisal of the quality and relevance of evidence. *Research Papers in Education*, *22*(2), 213–228. https://doi.org/10.1080/02671520701296189

Greiff, S., Molnár, G., Martin, R., Zimmermann, J., & Csapó, B. (2018). Students' exploration strategies in computer-simulated complex problem environments: A latent class approach. *Computers & Education, 126*, 248–263. https://doi.org/10.1016/j.compedu.2018.07.013

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

Griffin, P., McGaw, B., & Care, E. (2012). Assessment and teaching of 21st century skills. In *Assessment and teaching of 21st century skills*. Springer Science+Business Media. https://doi.org/10.1007/978-94-007-2324-5

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

Hamers, J. H. M., De Koning, E., & Sijtsma, K. (1998). Inductive reasoning in third grade: Intervention promises and constraints. *Contemporary Educational Psychology*, *23*(2), 132–148. https://doi.org/10.1006/ceps.1998.0966

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

Hanson, S. (2016). *The assessment of scientific reasoning skills of high school science students : A standardized assessment instrument* [Illinois State University]. https://ir.library.illinoisstate.edu/etd/506%0A

Hassler Hallstedt, M., & Ghaderi, A. (2018). Tablets instead of paper-based tests for young children? Comparability between paper and tablet versions of the mathematical Heidelberger Rechen Test 1-4. *Educational Assessment*, *23*(3), 195–210. https://doi.org/10.1080/10627197.2018.1488587

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Heckhausen, J., Wrosch, C., & Schulz, R. (2010). A motivational theory of life-span development. *Psychological Review*, *117*(1), 32–60. https://doi.org/10.1037/a0017668

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

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., & Liao, S. N. (2018). Predicting academic performance: a systematic literature review. *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, 175–199.

https://doi.org/10.1145/3293881.3295783

Hernesniemi, E., Räty, H., Kasanen, K., Cheng, X., Hong, J., & Kuittinen, M. (2020). Students' achievement motivation in Finnish and Chinese higher education and its relation to perceived teaching-learning environments. *Scandinavian Journal of Psychology, 61*(2), 204–217. https://doi.org/10.1111/sjop.12580

Hoang, K. M., Nguyen, H. T., & La, T. T. (2014). Parent and Teacher Communication: A Case Study in Vietnam. In W. J., X. B., & W. B. (Eds.), *Innovative Management in Information and Production* (pp. 305–313). Springer New York. https://doi.org/10.1007/978-1-4614-4857-0_33

Hoffman, B. (2015). The developmental trajectory of motivation. In *Motivation for Learning and Performance* (Vol. 5, pp. 79–106). Elsevier Inc. https://doi.org/10.1016/b978-0-12-800779-2.00004-x

Hooper, M., Mullis, I. V. S., & Martin, M. O. (2013). TIMSS 2015 context questionnaire framework. In I. V. S. Mullis & M. O. Martin (Eds.), *TIMSS 2015 Assessment Frameworks* (pp. 61–82). TIMSS & PIRLS International Study Center.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hu, W., Jia, X., Plucker, J. A., & Shan, X. (2016). Effects of a critical thinking skills program on the learning motivation of primary school students. *Roeper Review*, *38*(2), 70–83. https://doi.org/10.1080/02783193.2016.1150374

Hwang, M. H., Choi, H. C., Lee, A., Culver, J. D., & Hutchison, B. (2016). The relationship between self-efficacy and academic achievement: A 5-year panel analysis. *The Asia-Pacific Education Researcher*, *25*(1), 89–98. https://doi.org/10.1007/s40299-015-0236-3

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

Iliescu, D., & Greiff, S. (2019). The impact of technology on psychological testing in practice and policy: What will the future bring. *European Journal of Psychological Assessment*, *35*(2), 151–155. https://doi.org/10.1027/1015-5759/a000532

Jeotee, K. (2012). *Reasoning skills, problem solving ability and academic ability: implications for study programme and career choice in the context of higher education in Thailand (Unpublished doctoral dissertation)* [Durham University]. http://etheses.dur.ac.uk/3380

Józsa, K. (2014). Developing new scales for assessing English and German language mastery motivation. In J. Horváth & P. Medgyes (Eds.), *Studies in honour of Marianne Nikolov* (Issue July). Lingua Franca Csoport.

Józsa, K., Kis, N., & Barrett, K. C. (2019). Mastery motivation, parenting, and school achievement among Hungarian adolescents. *European Journal of Psychology of Education*, *34*(2), 317–339. https://doi.org/10.1007/s10212-018-0395-8

Józsa, K., Kis, N., & Huang, S. (2017). Mastery motivation in school subjects in Hungary and Taiwan. *Hungarian Educational Research Journal*, *7*(2), 158–177. https://doi.org/10.14413/HERJ/7/2/10

Kalinowski, S. T., & Willoughby, S. (2019). Development and validation of a scientific (formal) reasoning test for college students. *Journal of Research in Science Teaching*, *56*(9), 1269–1284. https://doi.org/10.1002/tea.21555

Kambeyo, L. (2018). *Assessing Namibian students ' abilities in scientific reasoning, scientific inquiry and inductive reasoning skills (Unpublished doctoral dissertation)*. University of Szeged.

Kambeyo, L., & Wu, H. (2018). Online assessment of students' inductive reasoning skills abilities in Oshana region , Namibia. *International Journal of Educational Sciences*, *21*, 1–12. https://doi.org/11.258359/KRE-86

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 198–211. https://doi.org/10.1080/0969594X.2015.1060192

Kaygısız, G. M., & Gürkan, B. (2018). Adaptation of Scientific Reasoning Scale into Turkish and Examination of its Psychometric Properties. *Educational Sciences: Theory & Practice*, *18*(3), 737–757. https://doi.org/10.12738/estp.2018.3.0175

Kim, D. H., & Huynh, H. (2010). Equivalence of paper-and-pencil and online administration modes of the statewide English test for students with and without disabilities. *Educational Assessment*, *15*(2), 107–121. https://doi.org/10.1080/10627197.2010.491066

King, R. B., & Ganotice, F. A. (2014). What's happening to our boys? A personal investment analysis of gender differences in student motivation. *Asia-Pacific Education Researcher*, *23*(1), 151–157. https://doi.org/10.1007/s40299-013-0127-4

Kinshuk, Lin, T., & Mcnab, P. (2006). Cognitive trait modelling: The case of inductive reasoning ability. *Innovations in Education and Teaching International*, *43*(2), 151–161. https://doi.org/10.1080/14703290600650442

Klauer, K. J., & Phye, G. D. (2008). Inductive reasoning: A training approach. *Review of*

Educational Research, *78*(1), 85–123. https://doi.org/10.3102/0034654307313402

Kniss AR, S. J. (2018). *Statistical Analysis of Agricultural Experiments using R*. https://rstats4ag.org

Köksal-tuncer, Ö., & Sodian, B. (2018). The development of scientific reasoning : Hypothesis testing and argumentation from evidence in young children. *Cognitive Development*, *48*(2018), 135–145. https://doi.org/10.1016/j.cogdev.2018.06.011

Korom, E., B. Németh, M., Pásztor, A., & Csapó, B. (2017). Relationship between scientific and inductive reasoning in grades 5 and 7. *Paper Presented at the 17th Biennial Conference of the European Association for Research on Learning and Instruction (EARLI)*.

Kriegbaum, K., Becker, N., & Spinath, B. (2018). The relative importance of intelligence and motivation as predictors of school achievement: A meta-analysis. *Educational Research Review*, *25*(October), 120–148. https://doi.org/10.1016/j.edurev.2018.10.001

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 drowth 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

Kyllonen, P., Hartman, R., Sprenger, A., Weeks, J., Bertling, M., McGrew, K., Kriz, S., Bertling, J., Fife, J., & Stankov, L. (2019). General fluid/inductive reasoning battery for a high-ability population. *Behavior Research Methods*, *51*(2), 507–522. https://doi.org/10.3758/s13428-018-1098-4

Lafraire, J., Rioux, C., Hamaoui, J., Girgis, H., Nguyen, S., & Thibaut, J.-P. (2020). Food as a borderline domain of knowledge: The development of domain-specific inductive reasoning strategies in young children. *Cognitive Development*, *56*(October 2019), 100946. https://doi.org/10.1016/j.cogdev.2020.100946

Lawson, A. (2000). Classroom test of scientific reasoning. In *Revised Edition: August 2000 by Anton E. Lawson, Arizona State University. Based on: Lawson, A.E. 1978. Development and validation of the classroom test of formal reasoning. Journal of Research in Science Teaching, 15(1): 11-24.* http://www.public.asu.edu/~anton1/AssessArticles/Assessments/Mathematics Assessments/Scientific Reasoning Test.pdf

Lawson, A. (2009). Basic inferences of scientific reasoning, argumentation, and discovery.

Science Education, *94*(2), 336–364. https://doi.org/10.1002/sce.20357

Lawson, A. E., Clark, B., Cramer-Meldrum, E., Falconer, K. A., Sequist, J. M., & Kwon, Y. J. (2000). Development of scientific reasoning in college biology: Do two levels of general hypothesis-testing skills exist? *Journal of Research in Science Teaching*, *37*(1), 81–101. https://doi.org/10.1002/(SICI)1098-2736(200001)37:1<81::AID-TEA6>3.0.CO;2-I

Lazonder, A. W., & Wiskerke-Drost, S. (2015). Advancing scientific reasoning in upper elementary classrooms: Direct instruction versus task structuring. *Journal of Science Education and Technology*, *24*(1), 69–77. https://doi.org/10.1007/s10956-014-9522-8

Leighton, J. P., & Sternberg, R. J. (2004). The nature of reasoning. In *The Nature of Reasoning* (pp. 3–11). Cambridge University Press. https://doi.org/10.1161/CIRCRESAHA.116.305012

Luecht, R. M., & Sireci, S. G. (2011). *A Review of Models for Computer-Based Testing. Research Report 2011-12*. College Board. www.collegeboard.org.

MacKinnon, D. P., Cheong, J., & Pirlott, A. G. (2012). Statistical mediation analysis. In H. Cooper (Ed.), *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.* (pp. 313–331). American Psychological Association. https://doi.org/10.1037/13620-018

Maftuh, B. (2011). Status of ICT integration in education in Southeast Asian countries. *Innovation of Classroom Teaching and Learning through Lesson Study*, *1*, 1–9.

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

Magis, David, & Facon, B. (2012). Angoff's delta method revisited: Improving DIF detection under small samples. *British Journal of Mathematical and Statistical Psychology*, *65*(2), 302–321. https://doi.org/10.1111/j.2044-8317.2011.02025.x

Magis, David, & Facon, B. (2014). deltaPlotR: An R package for differential item functioning fnalysis with Angoff's delta plot. *Journal of Statistical Software, Code Snippets, 59*(1).

Mahnane, L., & Hafidi, M. (2016). Automatic detection of learning styles based on dynamic Bayesian network in adaptive e-learning system. *International Journal of Innovation and Learning, 20*(3), 289. https://doi.org/10.1504/IJIL.2016.079067

Mohamadi, Z. (2018). Comparative effect of online summative and formative assessment on EFL student writing ability. *Studies in Educational Evaluation, 59*(July 2017), 29–40.

https://doi.org/10.1016/j.stueduc.2018.02.003

Martin, Michael O, Mullis, I. V. S., Hooper, M., Yin, L., Foy, P., & Palazzo, L. (2016). Creating and Interpreting the TIMSS 2015 context questionnaire scales. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015*. TIMSS & PIRLS International Study Center.

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

McCallum, R. S. (2017). Handbook of nonverbal assessment. In R. S. McCallum (Ed.), *Handbook of Nonverbal Assessment*. Springer International Publishing. https://doi.org/10.1007/978-3-319-50604-3

Mehraj, A. B. (2016). The predictive power of reasoning ability on academic achievement. *International Journal of Learning, Teaching and Educational Research*, *15*(1), 79–88.

Messick, S. (1995). Standards of validity and the validity of standards in performance asessment. *Educational Measurement: Issues and Practice*, *14*(4), 5–8. https://doi.org/10.1111/j.1745-3992.1995.tb00881.x

Mo, J. (2019). How is students’ motivation related to their performance and anxiety ? In *PISA in Focus, No. 92*. OECD Publishing. https://doi.org/https://doi.org/10.1787/d7c28431-en

MOET. (2009). *Tài liệu phân phối chương trình Vật lí THPT*.

MOET. (2017). *Chương trình giáo dục tổng thể [General education program]*.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Journal of Clinical Epidemiology*, *62*(10), 1006–1012. https://doi.org/10.1016/j.jclinepi.2009.06.005

Mollohan, K. N. (2015). *Epistemologies and scientific reasoning skills among undergraduate science students*. The Ohio State University.

Molnár, G. (2011). Playful fostering of 6- to 8-year-old students’ inductive reasoning. *Thinking Skills and Creativity*, *6*(2011), 91–99. https://doi.org/10.1016/j.tsc.2011.05.002

Molnár, G., & Csapó, B. (2011). Az 1–11 évfolyamot átfogó induktív gondolkodás kompetenciaskála készítése a valószínűségi tesztelmélet alkalmazásával [Constructing inductive reasoning competency scales for years 1–11 using IRT models]. *Magyar Pedagógia*, *111*(2), 127–140.

Molnár G, Csapó B. (2019). Making the Psychological Dimension of Learning Visible: Using Technology-Based Assessment to Monitor Students' Cognitive Development. *Frontiers in Psychology*. 10:1368. doi: 10.3389/fpsyg.2019.01368.

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

Morales M., with code developed by the R Development Core Team, with general advice from the R. listserv, & Murdoch., community and especially D. (2020). *sciplot: Scientific graphing functions for factorial designs* (R package version 1.1-1). https://cran.r-project.org/package=sciplot

Mousa, M., & Molnár, G. (2020). Computer-based training in Math improves inductive reasoning of 9- to 11-year-old children. *Thinking Skills and Creativity*, *37*(January), 100687. https://doi.org/10.1016/j.tsc.2020.100687

Muniz, M., Seabra, A. G., & Primi, R. (2012). Validity and reliability of the inductive reasoning test for children - IRTC. *Psicologia: Reflexão e Crítica*, *25*(2), 275–285. https://doi.org/10.1590/s0102-79722012000200009

Muthén, L. K., & Muthén, B. O. (2012). *Mplus statistical modeling software: Release 7.0*. CA: Muthén & Muthén.

Neumann, M. M., & Neumann, D. L. (2019). Validation of a touch screen tablet assessment of early literacy skills and a comparison with a traditional paper-based assessment. *International Journal of Research & Method in Education*, *42*(4), 385–398. https://doi.org/10.1080/1743727X.2018.1498078

Nhat, H. T., Lien, N. T., Tinh, N. T., Vu, N., Hang, T., & Trang, N. T. (2018). The development of critical thinking for students in Vietnamese schools : From policies to ractices. *American Journal of Educational Research*, *6*(5), 431–435. https://doi.org/10.12691/education-6-5-10

Nikolov, M., & Csapó, B. (2018). The relationships between 8th graders' L1 and L2 reading skills, inductive reasoning and socio-economic status in early English and German as a foreign language programs. *System*, *73*, 48–57. https://doi.org/10.1016/j.system.2017.11.001

Nikou, S. A., & Economides, A. A. (2018). Mobile-based assessment : A literature review of publications in major referred journals from 2009 to 2018. *Computers & Education*, *125*(2018), 101–119.

Nunes, T., & Csapó, B. (2011). Developing and assessing mathematical reasoning. In B.

Csapó & M. Szendrei (Eds.), *Framework for diagnostic assessment of mathematics* (pp. 15–76). Nemzeti Tankönyvkiadó.

Oblinger, D. G., & Oblinger, J. L. (2005). *Educating the next generation*. EDUCAUSE. https://www.educause.edu/ir/library/pdf/pub7101.pdf

OECD (2004). Problem solving for tomorrow's world: First measures of cross-curricular competencies from Pisa 2003. Paris: OECD.

OECD (2014). PISA 2012 Results Volume V): Creative problem solving: Students' skills in tackling real-life problems. Paris: OECD. http://dx.doi.org/10.1787/9789264208070-en

OECD. (2015). *Pisa 2015: results in focus* (Issue 853). OECD Publishing.

OECD. (2016). *PISA 2015 results (Volume I): Excellence and equity in Education*. OECD Publishing. https://doi.org/10.1787/9789264266490-en

OECD. (2017a). Parental involvement, student performance and satisfaction with life. In *PISA 2015 Results (Volume III): Students' Well-Being: Vol. III* (Issue Volume Iii, pp. 155–171). OECD Publishing. https://doi.org/10.1787/9789264273856-13-en

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

OECD (2017c). PISA 2015 Results (Volume V): Collaborative problem solving. Paris: OECD publishing. http://dx.doi.org/10.1787/9789264285521-en

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

Osborne, J. (2013). The 21st century challenge for science education: Assessing scientific reasoning. *Thinking Skills and Creativity*, *10*, 265–279. https://doi.org/10.1016/j.tsc.2013.07.006

Pásztor, A. (2016). *Technology-based assessment and development of inductive reasoning*. Ph.D. thesis. Doctoral School of Education, University of Szeged. doi:10.14232/phd.3191

Pásztor, A., Molnár, Gy., Korom, E., B. Németh, M., & Csapó, B. (2017). Online assessment of inductive reasoning and its predictive power on inquiry skills in science. *17th biennial conference of the European Association for Research on Learning and Instruction (EARLI). p509. Tampere, Finland, August 29-September 2, 2017*

Patrick, H., Mantzicopoulos, P., & Samarapungavan, A. (2009). Motivation for learning science in kindergarten: Is there a gender gap and does integrated inquiry and literacy instruction make a difference. *Journal of Research in Science Teaching*, *46*(2), 166–191. https://doi.org/10.1002/tea.20276

Peetsma, T., Hascher, T., van der Veen, I., & Roede, E. (2005). Relations between adolescents' self-evaluations, time perspectives, motivation for school and their achievement in different countries and at different ages. *European Journal of Psychology of Education*, *20*(3), 209–225. https://doi.org/10.1007/BF03173553

Pepper, D. (2011). Assessing key competences across the curriculum - and Europe. *European Journal of Education*, *46*(3), 335–353. https://doi.org/10.1111/j.1465-3435.2011.01484.x

Perret, P. (2015). Children 's inductive reasoning : Developmental and educational perspectives. *Journal of Cognitive Education and Psychology*, *14*(3), 389–408.

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*. http://www.thepiratesguidetor.com.

Phillips, N. (2017). *Yarrr: A companion to the e-Book "YaRrr!: The pirate's guide to R*. www.thepiratesguidetor.com

Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education: Theory, research, and applications*. Prentice Hall.

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

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*(3), 879–891. https://doi.org/10.3758/BRM.40.3.879

Preckel, F., Goetz, T., Pekrun, R., & Kleine, M. (2008). Gender Differences in Gifted and Average-Ability Students. *Gifted Child Quarterly*, *52*(2), 146–159. https://doi.org/10.1177/0016986208315834

R Core Team. (2019). R: A language and environment for statistical computing. In *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*. https://cran.r-project.org/package=BMA

Resing, W. C. M., Touw, K. W. J., Veerbeek, J., & Elliott, J. G. (2017). Progress in the inductive strategy-use of children from different ethnic backgrounds: a study employing dynamic testing. *Educational Psychology*, *37*(2), 173–191.

https://doi.org/10.1080/01443410.2016.1164300

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), e0146021. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0146021

Roberts, M. J., Welfare, H., Livermore, D. P., & Theadom, A. M. (2000). Context, visual salience, and inductive reasoning. *Thinking and Reasoning*, *6*(4), 349–374. https://doi.org/10.1080/135467800750038175

Robinson, K. A., Lee, Y., Bovee, E. A., Perez, T., Walton, S. P., Briedis, D., & Linnenbrink-Garcia, L. (2019). Motivation in transition: Development and roles of expectancy, task values, and costs in early college engineering. *Journal of Educational Psychology*, *111*(6), 1081–1102. https://doi.org/10.1037/edu0000331

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. http://www.jstatsoft.org/v48/i02/

Rudolph, J., Greiff, S., Strobel, A., & Preckel, F. (2018). Understanding the link between need for cognition and complex problem solving. *Contemporary Educational Psychology*, *55*(August), 53–62. https://doi.org/10.1016/j.cedpsych.2018.08.001

Şad, S. N. (2020). Does difficulty-based item order matter in multiple-choice exams? (Empirical evidence from university students). *Studies in Educational Evaluation*, *64*(September 2019), 100812. https://doi.org/10.1016/j.stueduc.2019.100812

Saleh, A. R., & Molnár, G. (2018). Inductive reasoning through the grades: case of Indonesia. *EDULEARN18 Proceedings : 10th International Conference on Education and New Learning Technologies*, 8790–8793.

Salihu, L., Aro, M., & Räsänen, P. (2018). Children with learning difficulties in mathematics: Relating mathematics skills and reading comprehension. *Issues in Educational Research*, *28*(4), 1024–1038.

Schroeders, U., & Wilhelm, O. (2010). Testing reasoning ability with handheld computers, notebooks, and paper and pencil. *European Journal of Psychological Assessment*, *26*(4), 284–292. https://doi.org/10.1027/1015-5759/a000038

Schwartz, S. J., & Waterman, A. S. (2006). Changing interests: A longitudinal study of intrinsic motivation for personally salient activities. *Journal of Research in Personality*, *40*(6), 1119–1136. https://doi.org/10.1016/j.jrp.2005.12.003

Schweizer, F., Wüstenberg, S., & Greiff, S. (2013). Validity of the MicroDYN approach:

Complex problem solving predicts school grades beyond working memory capacity. *Learning and Individual Differences*, *24*, 42–52. https://doi.org/10.1016/j.lindif.2012.12.011

Schwichow, M., Christoph, S., Boone, W. J., & Härtig, H. (2016). 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., Croker, S., Zimmerman, C., Höffler, T., & Härtig, H. (2016). 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., 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*, 101923. https://doi.org/10.1016/j.cedpsych.2020.101923

Shaakumeni, S. N., & Csapó, B. (2018). A cross-cultural validation of adapted questionnaire for assessing motivation to learn science. *African Journal of Research in Mathematics, Science and Technology Education*, *22*(3), 340–350. https://doi.org/10.1080/18117295.2018.1533157

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159–194. https://doi.org/10.1007/BF02294572

Sheard, M. K., & Chambers, B. (2014). A case of technology-enhanced formative assessment and achievement in primary grammar: How is quality assurance of formative assessment assured? *Studies in Educational Evaluation*, *43*, 14–23. https://doi.org/10.1016/j.stueduc.2014.02.001

Shute, V. J., & Rahimi, S. (2017). Review of computer-based assessment for learning in elementary and secondary education. *Journal of Computer Assisted Learning*, *33*, 1–19. https://doi.org/10.1111/jcal.12172

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

Soodmand Afshar, H., Rahimi, A., & Rahimi, M. (2014). Instrumental motivation, critical thinking, autonomy and academic achievement of iranian EFL learners. *Issues in Educational Research*, *24*(3), 281–298.

Spinath, B., Spinath, F. M., Harlaar, N., & Plomin, R. (2006). Predicting school achievement from general cognitive ability, self-perceived ability, and intrinsic value. *Intelligence*, *34*(4), 363–374. https://doi.org/10.1016/j.intell.2005.11.004

Stamovlasis, D., Tsitsipis, G., & Papageorgiou, G. (2010). The effect of three cognitive variables on students' understanding of the particulate nature of matter and its changes of state. *International Journal of Science Education*, *32*(08), 987–1016. https://doi.org/10.1080/09500690902893605

Staver, J. R. (1986). The effects of problem format, number of independent variables, and their interaction on student performance on a control of variables reasoning problem. *Journal of Research in Science Teaching*, *23*(6), 533–542. https://doi.org/10.1002/tea.3660230606

Steinmayr, R., & Spinath, B. (2009). The importance of motivation as a predictor of school achievement. *Learning and Individual Differences*, *19*(1), 80–90. https://doi.org/10.1016/j.lindif.2008.05.004

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

Sternberg, R. J. (1986). Toward a unified theory of human reasoning. *Intelligence*, *10*, 281–314.

Sternberg, R. J., & Sternberg, K. (2012). *Cognitive Psychology*. Cengage Learning products. https://doi.org/10.1039/ft9918702861

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

Stiller, J., Hartmann, S., Mathesius, S., Straube, P., Tiemann, R., Nordmeier, V., Krüger, D., & Upmeier zu Belzen, A. (2016). Assessing scientific reasoning: a comprehensive evaluation of item features that affect item difficulty. *Assessment and Evaluation in Higher Education*, *41*(5), 721–732. https://doi.org/10.1080/02602938.2016.1164830

Strobel, A., Behnke, A., Gärtner, A., & Strobel, A. (2019). The interplay of intelligence and need for cognition in predicting school grades: A retrospective study. *Personality and Individual Differences*, *144*(July 2018), 147–152. https://doi.org/10.1016/j.paid.2019.02.041

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

Tee, K. N., Leong, K. E., & Abdul Rahim, S. S. (2018). The mediating effects of critical thinking skills on motivation factors for mathematical reasoning ability. *Asia-Pacific Education Researcher*, *27*(5), 373–382. https://doi.org/10.1007/s40299-018-0396-z

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

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

TIMSS. (1997). *TIMSS Science Items : Released set for Population 2 (seventh and eighth grades)*. IEA TIMSS.

TIMSS. (2015). *Student questionnaire (separate Science Subjects)*.

Tsai, L.-T., Yang, C.-C., & Chang, Y.-J. (2015). Gender differences in factors affecting science performance of eighth grade Taiwan students. *The Asia-Pacific Education Researcher*, *24*(2), 445–456. https://doi.org/10.1007/s40299-014-0196-z

Tuan, H. L., Chin, C. C., & Shieh, S. H. (2005). The development of a questionnaire to measure students' motivation towards science learning. *International Journal of Science Education*, *27*(6), 639–654. https://doi.org/10.1080/0950069042000323737

Tunteler, E., Pronk, C. M. E., & Resing, W. C. M. (2008). Inter- and intra-individual variability in the process of change in the use of analogical strategies to solve geometric tasks in children: A microgenetic analysis. *Learning and Individual Differences*, *18*(1), 44–60. https://doi.org/10.1016/j.lindif.2007.07.007

Tzuriel, D., & George, T. (2009). Improvement of analogical reasoning and academic achievements by the analogical reasoning programme (ARP). *Educational and Child Psychology*, *26*(3), 71–94.

UNESCO. (2011). *World data on Education*.

http://www.ibe.unesco.org/fileadmin/user_upload/Publications/WDE/2010/pdf-versions/Viet_Nam.pdf

Vainikainen, M.-P. (2014). *Finnish primary school pupils' performance in learning to learn assessments: A longitudinal perspective on educational equity*.

Vainikainen, M.-P., Hautamäki, J., Hotulainen, R., & Kupiainen, S. (2015). General and specific thinking skills and schooling: Preparing the mind to new learning. *Thinking Skills and Creativity*, *18*(2015), 53–64. https://doi.org/http://dx.doi.org/10.1016/j.tsc.2015.04.006

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. (n.d.). Measuring inductive reasoning in school contexts : a review of instruments and predictors. *Int. J. Innovation and Learning (in Press)*.

Van Vo, D., & Csapó, B. (2020). Development of inductive reasoning in students across school grade levels. *Thinking Skills and Creativity*, *37*(2020), 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

Venville, G., & Oliver, M. (2015). The impact of a cognitive acceleration programme in science on students in an academically selective high school. *Thinking Skills and Creativity*, *15*(2015), 48–60. https://doi.org/10.1016/j.tsc.2014.11.004

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

Vogelaar, B., Sweijen, S. W., & Resing, W. C. M. (2019). Gifted and average-ability children's potential for solving analogy items. *Journal of Intelligence*, *7*(3). https://doi.org/10.3390/jintelligence7030019

Voogt, J., & Roblin, N. P. (2012). A comparative analysis of international frameworks for 21stcentury competences: Implications for national curriculum policies. *Journal of Curriculum Studies*, *44*(3), 299–321. https://doi.org/10.1080/00220272.2012.668938

Wang, K. (2008). *Investigating the domain of geometric inductive reasoning problems: a structural equation modeling analysis (Doctoral dissertation)* (Issue April). Brigham Young University.

Waschl, N. A., Nettelbeck, T., Jackson, S. A., & Burns, N. R. (2016). Dimensionality of the Raven's Advanced Progressive Matrices: Sex differences and visuospatial ability. *Personality and Individual Differences*, *100*, 157–166. https://doi.org/10.1016/j.paid.2015.12.008

Waschl, N., & Burns, N. R. (2020). Sex differences in inductive reasoning: A research synthesis using meta-analytic techniques. *Personality and Individual Differences*, *164*(February), 109959. https://doi.org/10.1016/j.paid.2020.109959

Wesiak, G. (2003). *Ordering inductive reasoning tests for adaptive knowledge assessments: An application of surmise relations between tests (Unpublished doctoral dissertation)* (Issue June). University of Graz.

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

Williams, J. E., & McCord, D. M. (2006). Equivalence of standard and computerized versions of the Raven Progressive Matrices Test. *Computers in Human Behavior*, *22*(5), 791–800. https://doi.org/10.1016/j.chb.2004.03.005

Williamson, K. C., Williamson, V. M., & Hinze, S. R. (2017). Administering Spatial and Cognitive Instruments In-class and On-line: Are These Equivalent? *Journal of Science Education and Technology*, *26*(1), 12–23. https://doi.org/10.1007/s10956-016-9645-1

Wilson, M., De Boeck, P., & Carstensen, C. H. (2008). Explanatory item response models: A brief introduction. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of Competencies in Educational Contexts* (pp. 83–110). Hogrefe & Huber.

Wise, S. L., & Kuhfeld, M. R. (2021). Using retest data to evaluate and improve effort-moderated scoring. *Journal of Educational Measurement*, *58*(1), 130–149. https://doi.org/10.1111/jedm.12275

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, H., & Molnár, G. (2018). Interactive problem solving: Assessment and relations to combinatorial and inductive reasoning. *Journal of Psychological and Educational Research*, *26*(1), 90–105.

Yanto, B. E., Subali, B., & Suyanto, S. (2019). Measurement instrument of scientific reasoning test for Biology education students. *Distance Education*, *12*(1), 1694-609X.

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. CA: University of California.

Zeyer, A. (2010). Motivation to learn science and cognitive style. *Eurasia Journal of Mathematics, Science and Technology Education*, *6*(2), 123–130. https://doi.org/10.12973/ejmste/75233

Zeyer, A., & Wolf, S. (2010). Is there a relationship between brain type, sex and motivation to learn science? *International Journal of Science Education*, *32*(16), 2217–2233. https://doi.org/10.1080/09500690903585184

Zhang, F., & Bae, C. L. (2020). Motivational factors that influence student science achievement: a systematic literature review of TIMSS studies. *International Journal of Science Education*, *42*(17), 2921–2944. https://doi.org/10.1080/09500693.2020.1843083

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. (2000). The development of scientific reasoning skills. *Developmental Review*, *20*(1), 99–149. https://doi.org/10.1006/drev.1999.0497

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

Appendix A. The main properties of the studies selected (N= 63)

Reference	Country/ Sample	Level	Main Instrument	Domain	Mode
Csapó (1997)	Hungary 2424	I, II, III	IR, DR, RPM, science test	g	pp
Valanides (1997)	Cyprus 469	III	SR	g	pp
Hamers et al. (1998)	Netherlands 350	I	RPM	g	pp
Kwon & Lawson (2000)	Korea 210	II, III	DR (Wisconsin Card Sorting Test), LCTSR, physics	g, s	na
Roberts et al. (2000)	Britain 445	I	IR (analogies, geometric matrices)	g	pp
Gerber et al. (2001)	United States 550	II, III	LCTSR	g	
Wesiak (2003)	Austria 122	III, IV	IR	g	pp
Williams & McCord (2006)	United States 50	IV	RPM	g	mix
Tekkaya & Yenilmez (2006)	Turkey 117	II	SR, LAQ		pp
Boujaoude et al. (2007)	Lebanon 68	III	SR, LQ, chemistry test	g, s	pp
Tunteler et al. (2008)	Netherlands 36	I	IR (analogies)	g	pp
Wang (2008)	China 334	II	IR	g	on
Tzuriel & George (2009)	na 53	I	IR (analogies) Basic Math, Reading	g	pp
Bao et al. (2009)	United States, China 3000	IV	LCTSR, physics tests	g, s	na
Dilivio (2009)	United States 203	IV	IR (analogies)	g	on
Schroeders & Wilhelm (2010)	Germany 157	III	IR, SR(verbal, figural, and numerical)	g	mi
Stamovlasis et al. (2010)	Greece 329	II	LCTSR, physics test	g, s	pp
Boroş & Sas (2011)	Romania 30	I	IR	g	pp
Molnár (2011)	Hungary 252	I	IR, Klauer's program	g	pp
Barkl et al. (2012)	Australia 48	I	IR, DR	g, s	pp
Jeotee (2012)	Thailand 333	IV	IR, DR, PS	g	pp

Muniz et al. (2012)	Brazil 417	I	IR	g	pp
Chuang & She (2013)	na 190	I	LCSRT IR (Analogical reasoning) physics test	g, s	pp
Stevenson et al. (2013)	Netherlands 252	I	IR (Dynamic Analogical) - Working Memory - Math	g	on
Díaz-Morales & Escribano (2013)	Spain 887	II, III	IR, Morningness–Eveningness Scale, Sleep Habits Survey	g	na
Molnár et al. (2013)	Hungary 2769	I, II, III	IR, PS	g	on
Csapó et al. (2014)	Hungary 435	I	IR, speech sound discrimination test	g	on
Mayer et al. (2014)	Germany 155	I	SR, IR, PS, reading, spatial abilities	g	pp
Piraksa et al. (2014)	Thailand 400	III	LCTSR	g	pp
Lazonder & Wiskerke-Drost (2015)	Netherlands 55	I	SR (CVS)	s	pp
Mollohan (2015)	United States 987	IV	LCTSR, Learning Attitudes about Science	g	pp
Tairab (2015)	United Arab Emirates 128	II, III	SR (CVS)	g	pp
Thuneberg et al. (2015)	Finland 769	IV	SR (CVS)	g	pp
Vainikainen et al. (2015)	Finland 1543	I, II	IR, tests	g, s	mi
Venville & Oliver (2015)	Australia 582	II	IR, tests (math, science, writing)	g, s	pp
Blum et al. (2016)	Argentina: 422 Germany: 84	IV	IR	g	mi
Ding et al. (2016)	China 1637	IV	LCTSR	g	pp
Zhou, Han, Koenig, Raplinger, & Pi (2016)	United States: 189 China: 314	III, IV	SR (COV) in Physics	s	pp
Waschl, Nettelbeck, Jackson, & Burns (2016)	Australia 2105	IV	RPM	g	on
Ariës, Ghysels, Groot, & Brink (2016)	Netherlands 81	III	DR, WM function test	g	na
Mehraj (2016)	India 598	III	IR, DR, SR	g	na
Schwichow, Christoph, Boone, & Härtig (2016)	Germany 386	II	SR (COV)	g, s	pp

Al-Balushi, Al-Musawi, Ambusaidi, & Al-Hajri (2017)	Oman 60	III	A spatial ability SR(Chemistry)	s	on
Edelsbrunner (2017)	Switzerland 1809	I, II	SR (COV), BQ	s	pp
Resing et al. (2017)	Netherlands 116	I	IR(series)	g	mi
Ding (2018)	China 2669	I,II,III, IV	LCTSR	g	pp
Bao et al. (2018)	United States 1576	IV	LCTSR	g	na
Hejnová et al. (2018)	Czech Republic 23	III	LCTSR Culture of problem solving	g	pp
Kambeyo (2018)	Namibia 582	III	IR, SI, LCTSR, BQ	g	pp
Kambeyo & Wu (2018)	Namibia: 621 China: 25	II	IR, SR, MQ	g	on
Kaygısız & Gürkan (2018)	Turkey 947	IV	SR, LCTSR	g	na
Kyllonen et al. (2018)	United States 2000	IV	IR, BQ, MQ	g	on
Nikolov & Csapó, 2018	Hungary 1943	II	IR, reading	g	mix
Rudolph et al. (2018)	Germany 474	II	IR, PS	g, s	mi
Salihu et al. (2018)	Kosovo 233	I	RPM, math, reading tests	g	pp
Stender et al. (2018)	Germany 189	II, III	SR (physics), IR (figural analogies), reading and physics tests	g, s	pp
Wu & Molnár (2018)	China 187	II	IR, PS	g	on
Vogelaar et al. (2019)	Netherlands 74	I	RPM IR (Dynamic Analogical)	g	pp
Boğar (2019)	Turkey 60	II	SR (physics)	s	pp
Strobel et al. (2019)	Germany 290	IV	the Intelligence Structure Test 2000	g	on
Yanto et al. (2019)	Indonesia 100	IV	SR (biology)	s	na
Mousa & Molnár (2020)	Palestine 236	I	IR	g, s	on
Van Vo & Csapó (2020)	Vietnam 701	I, II, III	IR, BQ	g	mi

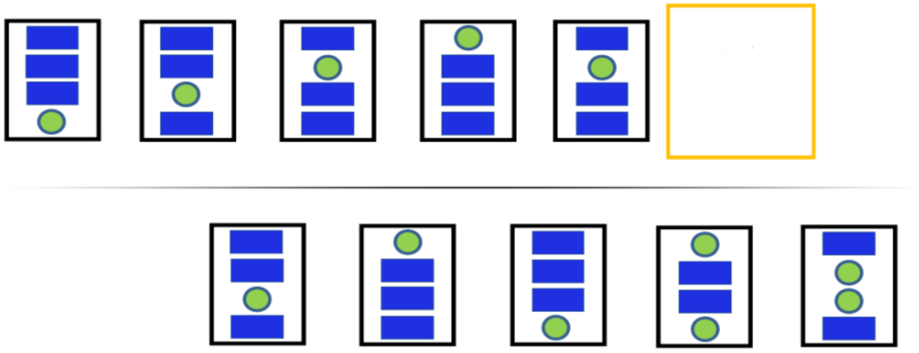
Note: I: 1st grade – 5th grade; II: 6th - 9th grades, III: 10th - 12th grades, IV: higher education; IR: inductive reasoning tasks; DR: deductive reasoning tasks; RPM: Raven's Standard Progressive Matrices; LAQ: The Learning Approach Questionnaire; LCTSR: The Lawson Classroom Test of Scientific Reasoning; LQ: Learning orientation questionnaire; PS: problem-solving test; MQ: Motivation Questionnaire; BQ: Background questionnaire; g: domain-general, s: domain-specific; pp: paper and pencil; on: online; mi: paper-and-pencil and online, na: unreported.

Appendix B. The inductive reasoning test (Vietnamese version)

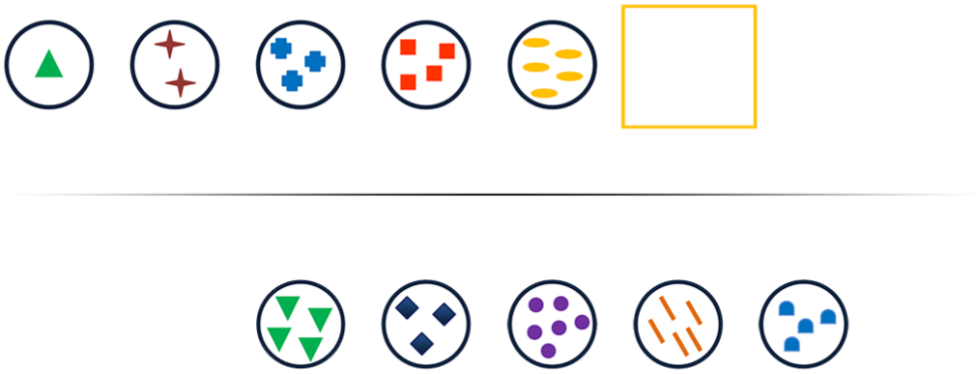
Yêu cầu chung từ Câu 1 đến Câu 8:

*Chọn **một hình** bên dưới để đặt vào khung hình còn trống của dãy hình phía trên sao cho phù hợp nhất.*

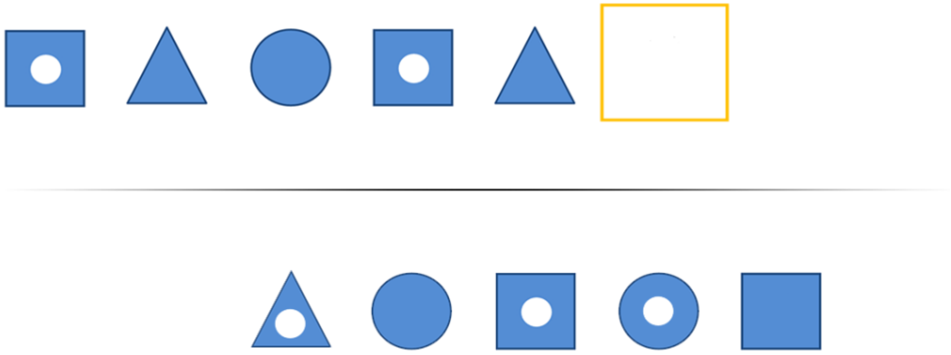
Câu 1 (IR_FS01)



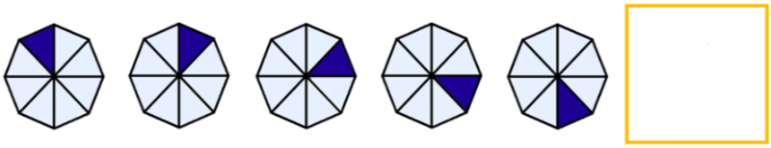
Câu 2 (IR_FS02)

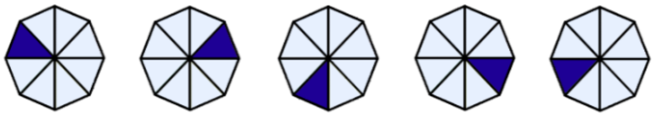


Câu 3 (IR_FS03)

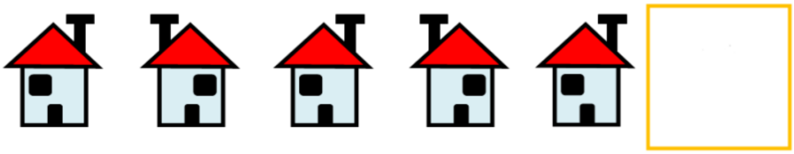


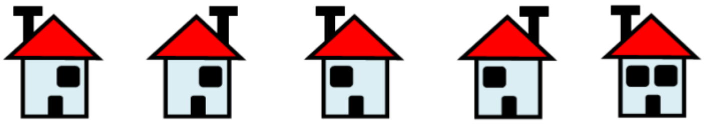
Câu 4 (IR_FS04)



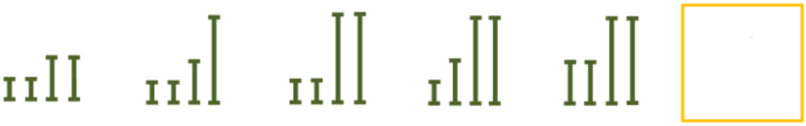


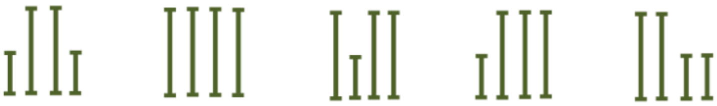
Câu 5 (IR_FS05)



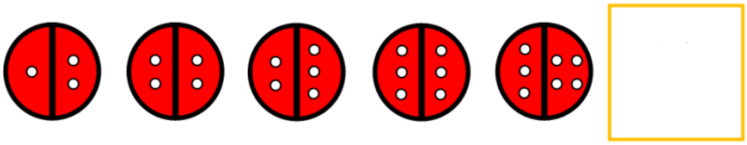


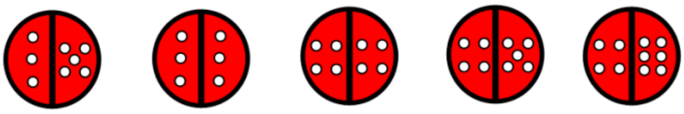
Câu 6 (IR_FS06)



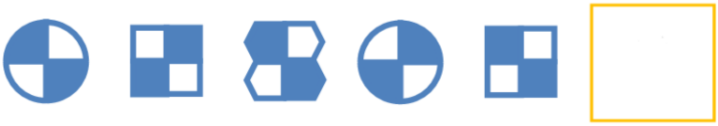


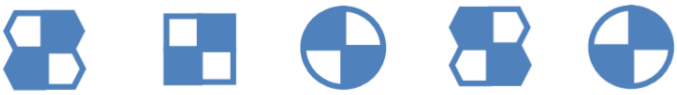
Câu 7 (IR_FS08)





Câu 8 (IR_FS10)

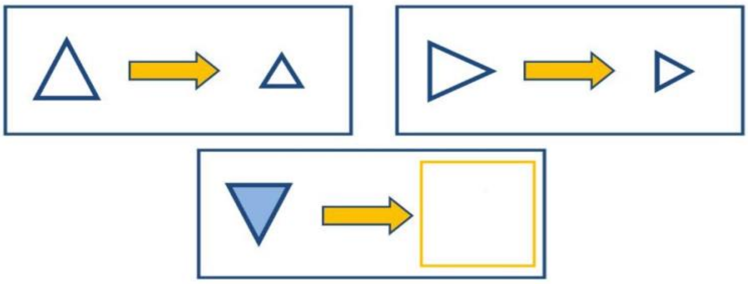


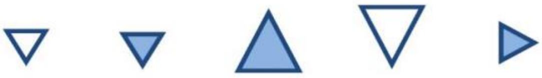


Yêu cầu chung từ Câu 9 đến Câu 16:

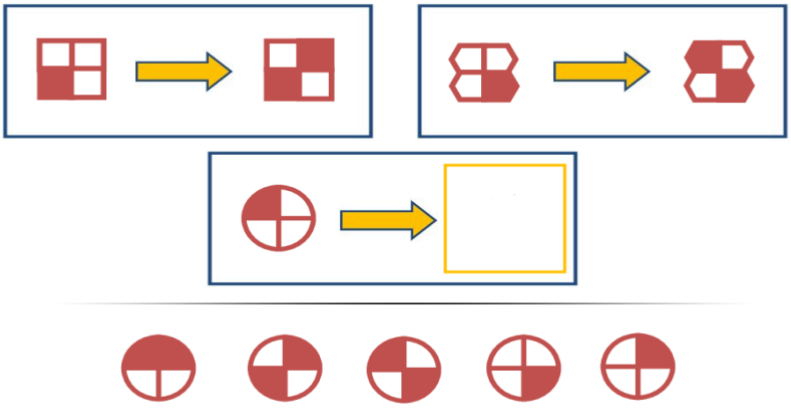
Hãy tìm quy luật từ hai cặp hình đã cho để xác định hình còn lại của cặp thứ ba. Chọn một hình bên dưới để đặt vào khung hình còn trống của dãy hình bên trên sao cho phù hợp quy luật nhất.

Câu 9 (IR_FA01)

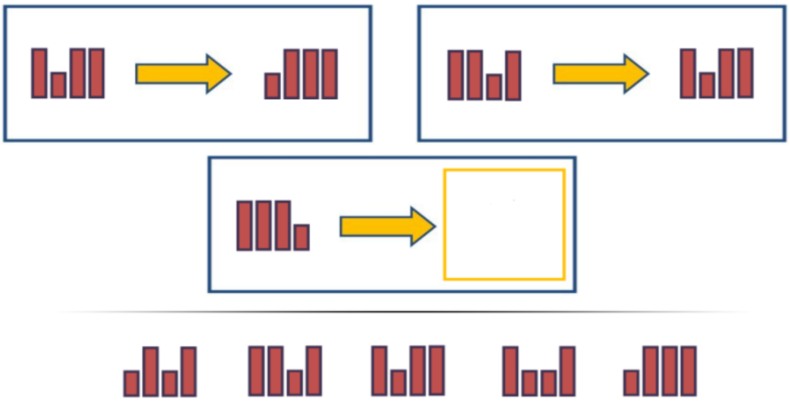




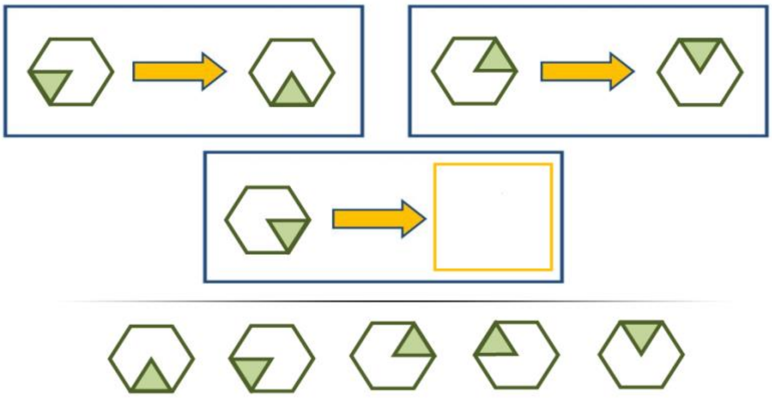
Câu 10 (IR_FA03)



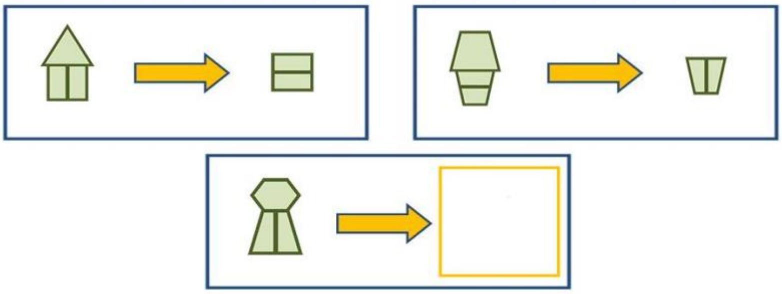
Câu 11 (IR_FA04)

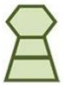
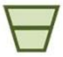
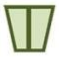
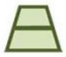
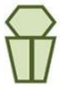


Câu 12 (IR_FA05)

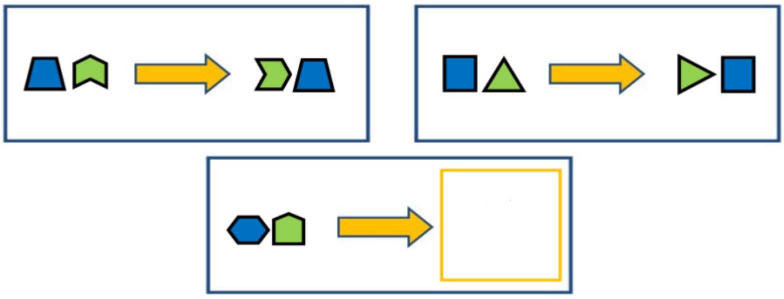


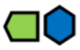
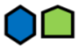
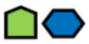
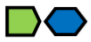
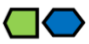
Câu 13 (IR_FA06)



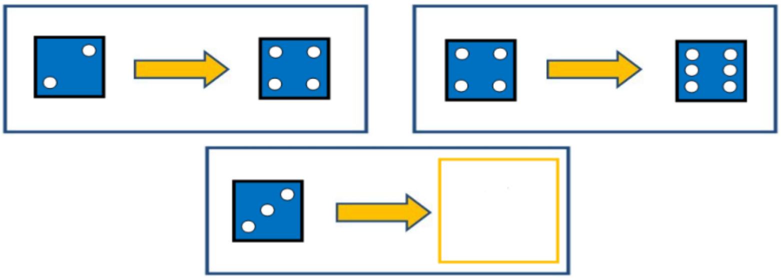






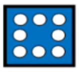
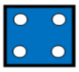
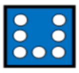
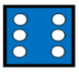
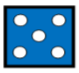
Câu 14 (IR_FA07)



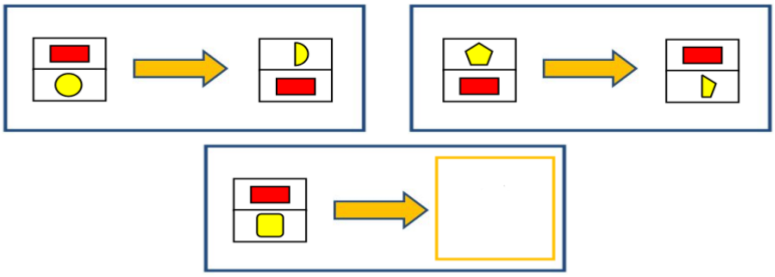






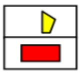
Câu 15 (IR_FA08)

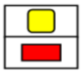


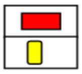






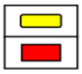
Câu 16 (IR_FA09)

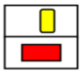












Yêu cầu chung từ Câu 17 đến Câu 24:

Hãy tìm quy luật từ hai cặp số đã cho để xác định số còn lại trong cặp thứ ba.

Chọn một số trong dãy số đã cho bên dưới để đặt vào khung còn trống của dãy số bên trên sao cho phù hợp quy luật nhất.

Câu 17 (IR_NA01)

20 → 32

8 → 20

11 → □

20 23 22 33 12

Câu 18 (IR_NA02)

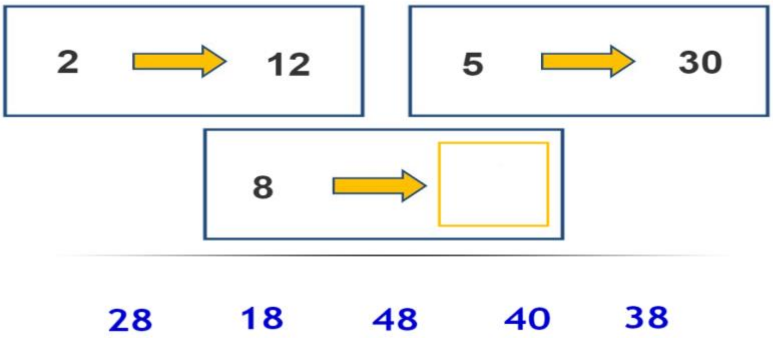
6 → 21

16 → 31

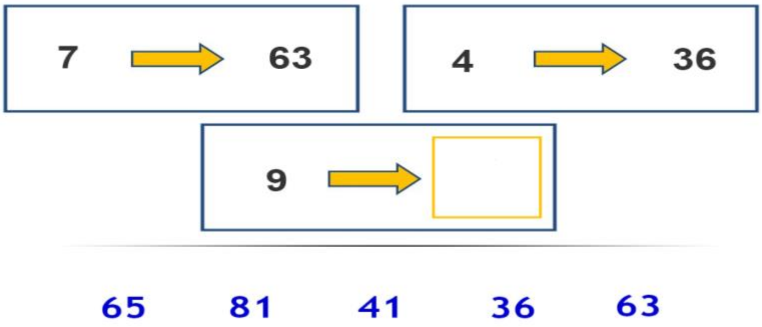
21 → □

31 41 6 15 36

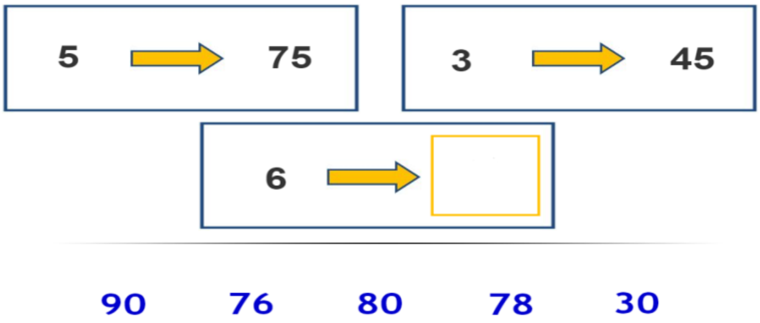
Câu 19 (IR_NA03)



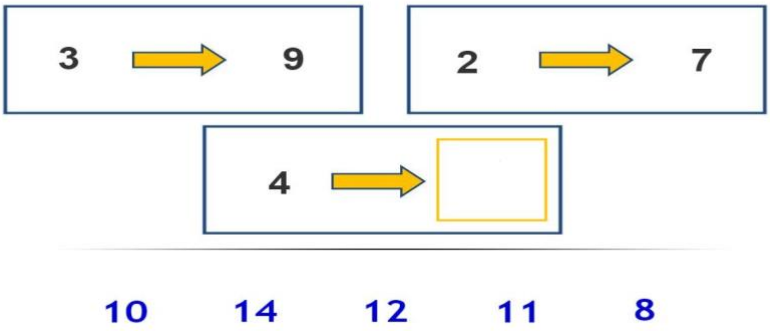
Câu 20 (IR_NA04)



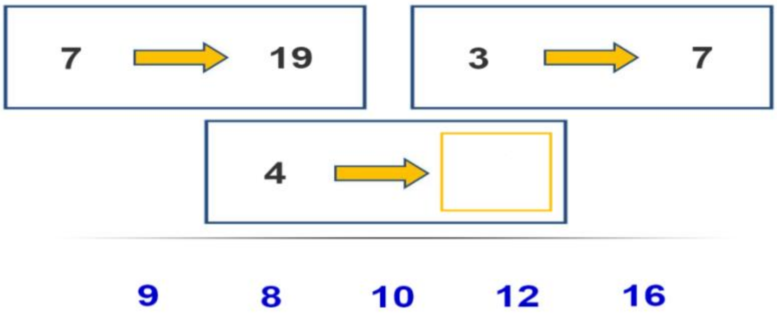
Câu 21 (IR_NA05)



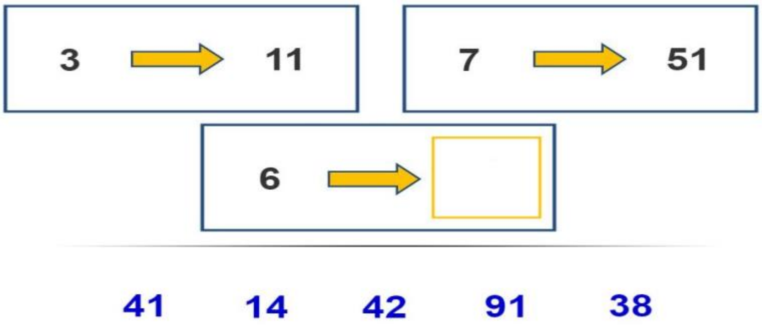
Câu 22 (IR_NA06)



Câu 23 (IR_NA08



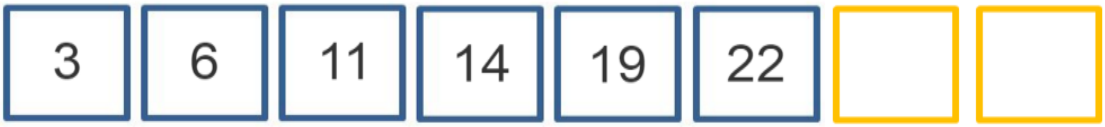
Câu 24 (IR_NA10)



Yêu cầu chung từ Câu 25 đến Câu 32:

Trong phần này, ta tìm quy luật từ dãy số đã cho để xác định hai số tiếp theo trong dãy. Chọn hai số trong dãy số đã cho bên dưới để đặt vào hai khung còn trống của dãy số bên trên sao cho phù hợp quy luật nhất.

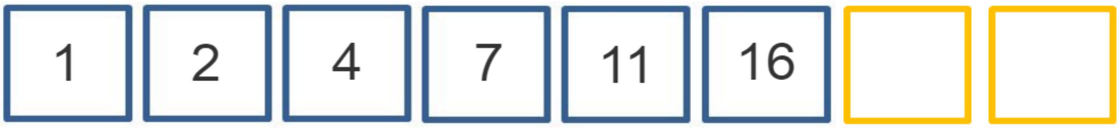
Câu 25 (IR_NS03)



3	6	11	14	19	22		

26 28 30 25 32 27 29

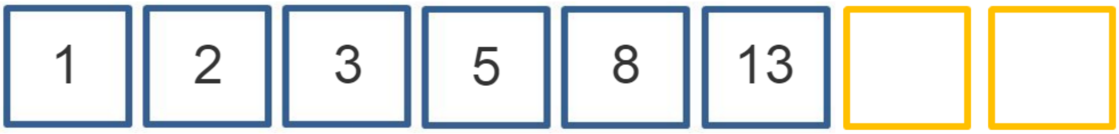
Câu 25 (IR_NS04)



1	2	4	7	11	16		

26 17 20 19 22 21 29

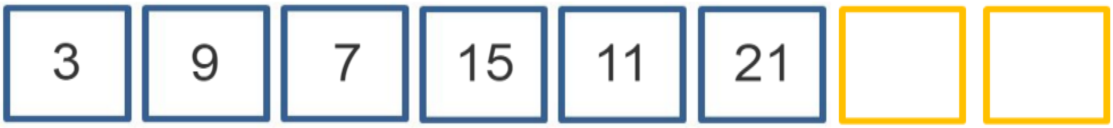
Câu 27 (IR_NS05)



1	2	3	5	8	13		

14 21 18 34 19 23 26

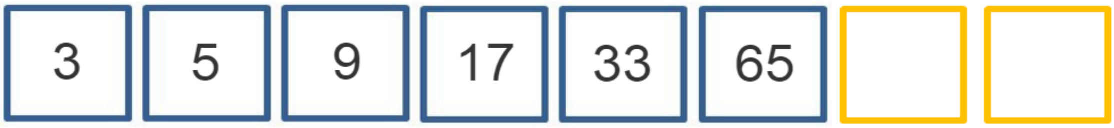
Câu 28 (IR_NS06)



3	9	7	15	11	21		

16 25 31 41 27 19 15

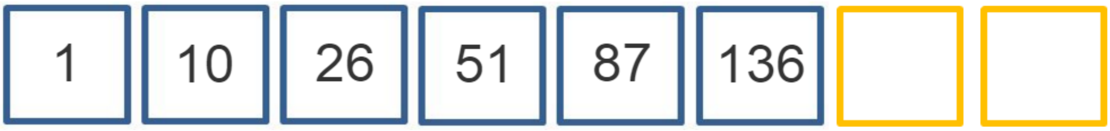
Câu 29 (IR_NS07)



3	5	9	17	33	65		

99 61 109 257 129 57 217

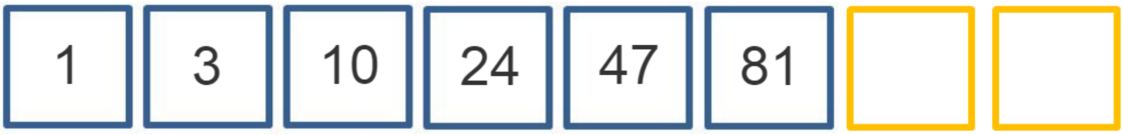
Câu 30 (IR_NS08)



1	10	26	51	87	136		

145 161 146 162 151 200 281

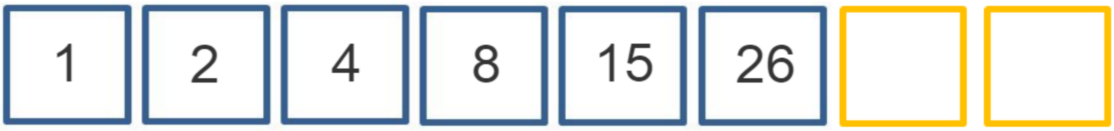
Câu 31 (IR_NS09)



1	3	10	24	47	81		

109 84 128 94 83 90 190

Câu 32 (IR_NS10)



1	2	4	8	15	26		

29 42 64 37 48 27 45

Appendix C. The scientific reasoning test (Vietnamese version)

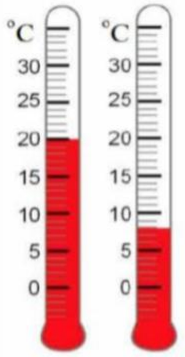
Câu 1(S01). Chúng ta cho sữa từ một ly thuỷ tinh sang một chén sành. Kết luận nào sau đây là **đúng**?

A. Cả thể tích và hình dạng của sữa thay đổi.
B. Chỉ có thể tích thay đổi nhưng hình dạng của sữa không đổi.
C. Chỉ hình dạng sữa thay đổi còn thể tích thì không.
D. Cả hình dạng và thể tích của sữa thay đổi.

Câu 2(S02). Chúng ta chuyển viên bi từ cốc thuỷ tinh nhỏ sang cốc thuỷ tinh lớn hơn. Kết luận nào sau đây **đúng**?

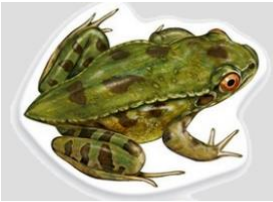
A. Cả thể tích và hình dạng viên bi thay đổi.
B. Chỉ có thể tích viên bi thay đổi còn hình dạng thì không.
C. Chỉ có hình dạng viên bi thay đổi còn thể tích thì không.
D. Cả hình dạng và thể tích của viên bi đều không thay đổi.

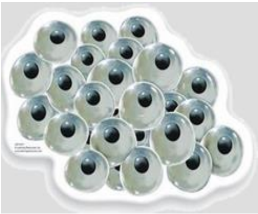
Câu 3(S03). Bạn Trâm muốn đo nhiệt độ ngoài trời vào buổi sáng. Bạn ấy mang nhiệt kế từ trong nhà ra ngoài trời. Hình bên thể hiện sự thay đổi xảy ra khi nhiệt kế đã đặt bên ngoài được vài phút. Những thuộc tính nào của chất lỏng trong nhiệt kế **không** thay đổi?

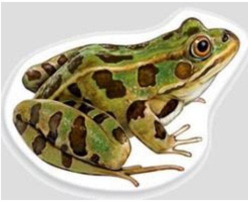


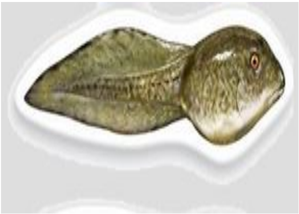
A. Khối lượng. **B.** Thể tích.
C. Nhiệt độ. **D.** Khối lượng riêng.

Câu 4(S04). Trình tự nào sau đây của các hình bên dưới là phù hợp nhất?


(I)

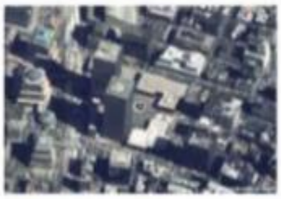

(II)


(III)

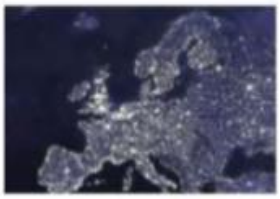

(IV)

A. (I) – (II) – (III) – (IV) **B.** (II) – (IV) – (I) – (III)
C. (II) – (IV) – (III) – (I) **D.** (IV) – (II) – (III) – (I)

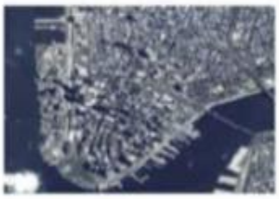
Câu 5(S05). Trình tự nào sau đây của các hình bên dưới là phù hợp nhất?



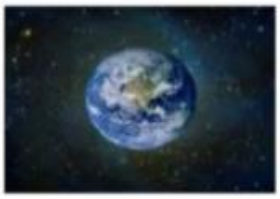
(I)



(II)



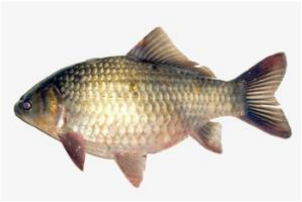
(III)



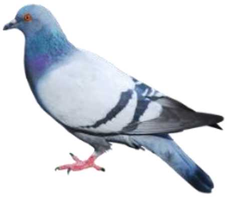
(IV)

A. (I) – (II) – (III) – (IV) **B.** (IV) – (III) – (II) – (I)
C. (II) – (IV) – (III) – (I) **D.** (IV) – (II) – (III) – (I)

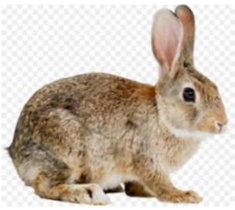
Câu 6(S06). Thứ tự nào sao đây của các hình là phù hợp nhất?



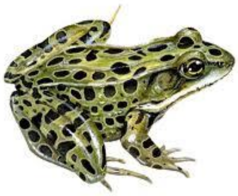
(I)



(II)



(III)



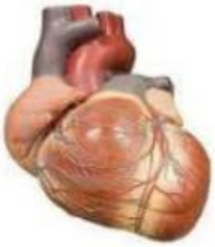
(IV)

A. (I) – (II) – (III) – (IV) **B.** (IV) – (III) – (II) – (I)
C. (I) – (IV) – (II) – (III) **D.** (IV) – (II) – (III) – (I)

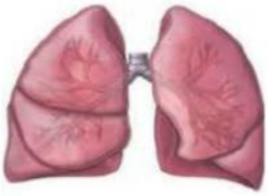
Câu 7(S07). Thiên thể nào sau đây **không** thuộc nhóm này?
A. Trái Đất **B.** Mặt Trăng **C.** Sao hoả **D.** Sao kim

Câu 8(S08). Phát biểu nào sau đây **không** thuộc nhóm này?
A. Một vũng nước bốc hơi . **B.** Băng đóng trên nhánh cây.
C. Lũ đang về trên sông. **D.** Giọt sương rơi trên chiếc lá.

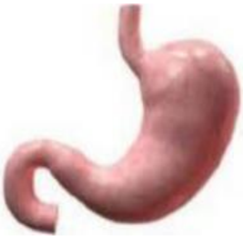
Câu 9(S09). Hình nào sau đây **không** thuộc nhóm này?



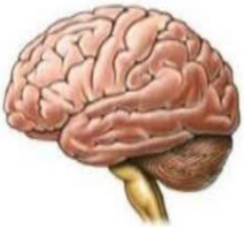
A



B



C



D

Câu 10(S10). Từ logic ở hai khung đầu, hãy suy luận tìm phương án hợp lí đặt vào khung có dấu chấm hỏi?

Kim loại	⇨	**Nhựa**		**Rắn**	⇨	**?**

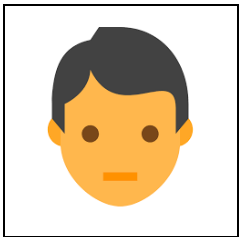
A. Sắt **B.** Lỏng **C.** Trạng thái vật chất. **D.** Gỗ

Câu 11(S11). Từ logic ở hai khung đầu, hãy suy luận tìm từ hợp lí đặt vào khung có dấu chấm hỏi?

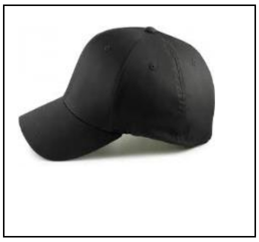
Mặt trời	⇨	**Trái đất**		**Nam châm**	⇨	**?**

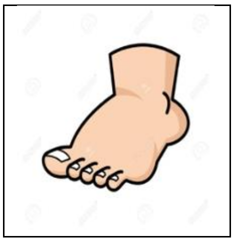
A. Cái thìa gỗ **B.** Cây kim **C.** Con dao bằng gốm **D.** Máy tính

Câu 12(S12). Từ logic ở hai khung đầu, hãy suy luận chọn hình hợp lí đặt vào khung có dấu chấm hỏi?



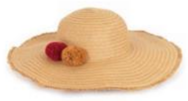
⇨



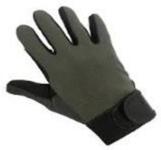


⇨

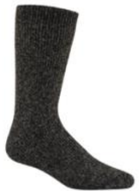
?



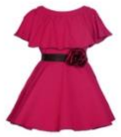
A



B



C



D

Câu 13(S13). Minh, Trâm và Dũng đánh dấu một thành phố làm điểm du lịch trên bản đồ. Chúng có khoảng cách 10 cm so với thủ đô nhưng với bản đồ có tỉ lệ khác nhau (xem bảng bên dưới). Hãy giúp từng bạn chọn loại phương tiện giao thông phù hợp cho kỳ nghỉ của họ.

Tên học sinh	Tỉ lệ bản đồ
Minh	1:1.500.000
Trâm	1:40.000
Dũng	1:11.600.000

A. Minh: Xe đạp, **Trâm**: xe ô-tô, **Dũng**: máy bay.
B. Minh: Xe ô-tô, **Trâm**: xe đạp, **Dũng**: máy bay.
C. Minh: máy bay, **Trâm**: xe ô-tô, **Dũng**: xe đạp.
D. Minh: Xe ô-tô, **Trâm**: máy bay, **Dũng**: xe đạp.

Câu 14(S14). Một chú Kăng-ga-ru (chuột túi) chạy nhanh gấp hai lần một chú voi. Phát biểu nào sao đây **đúng**?

A. Cùng một thời gian, chú Kăng-ga-ru có thể chạy gấp đôi đoạn đường so với chú voi.
B. Trên cùng một đoạn đường, chú Kăng-ga-ru có thể chạy với thời gian gấp đôi chú voi.
C. Cùng một thời gian, chú voi có thể chạy gấp đôi đoạn đường so với chú Kăng-ga-ru.
D. Cùng một khoản cách, chú voi có thể chạy với thời gian bằng một nữa so với chú Kăng-ga-ru.

Câu 15(S15). Chu kỳ tự quay của Trái Đất là một ngày, nghĩa là 24 giờ và Trái Đất tự quay được 360 độ xung quanh trục của nó. Hỏi trong 8 giờ, Trái đất sẽ thực hiện được:

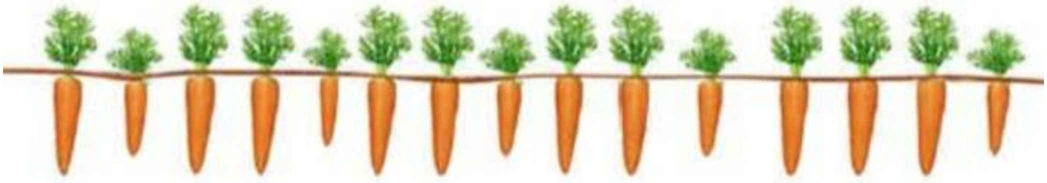
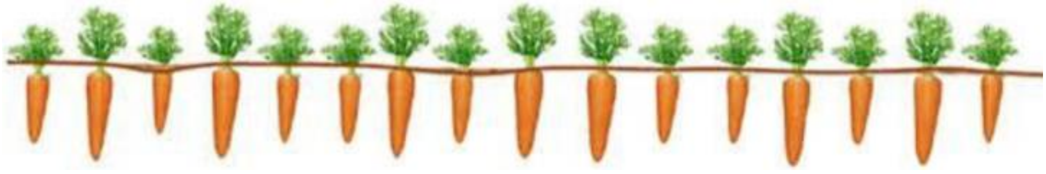
A. 1/3 ngày, tự quay 120 độ. **B.** 1/2 ngày, tự quay 180 độ.
C. 1/6 ngày, tự quay 36 độ. **D.** 1/4 ngày, tự quay 90 độ.

Câu 16(S16). Trong trường có một bài kiểm tra sức khoẻ học sinh. Người ta tìm ra rằng trong lớp có một số học sinh bị béo phì. Số liệu được ghi nhận từ ba lớp được thể hiện trong bảng bên dưới. Liệu rằng giới tính (nam/nữ) có ảnh hưởng đến tỉ lệ béo phì hay không?

Giới tính	Số lượng	
	Béo phì	Bình thường
Nam	8	38
Nữ	11	43

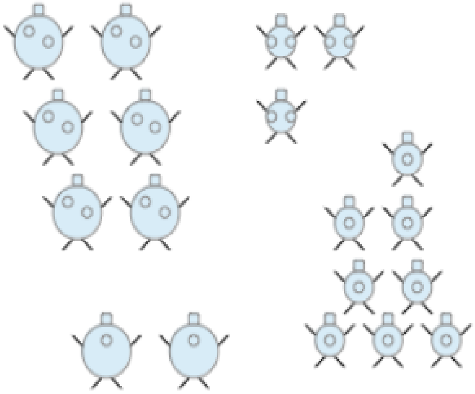
A. Có. **B.** Không. **C.** Không thể xác định

Câu 17(S17). Có sự ảnh hưởng của phân bón lên kích thước của Cà-rốt hay không?

Có sử dụng phân bón	
Không sử dụng phân bón	

A. Có. **B.** Không. **C.** Không thể xác định.

Câu 18(S18). Người ta thu được một lượng rùa biển. Chúng có đặt điểm: tất cả rùa có thể có một đốm hoặc hai đốm trên lưng (xem hình bên).



Em có thể kết luận gì về quan hệ giữa kích thước và số đốm trên lưng rùa trong mẫu này?

A. Có mối quan hệ rất rõ giữa kích thước và số đốm trên lưng rùa.

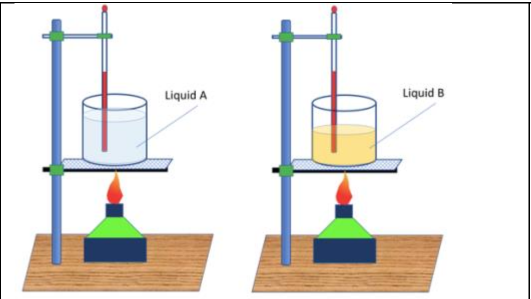
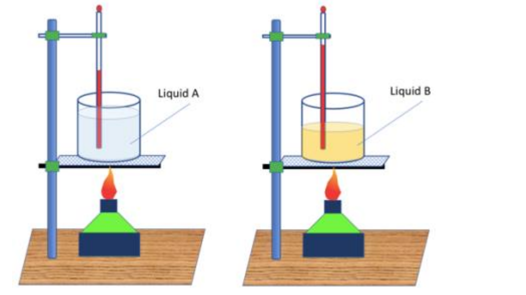
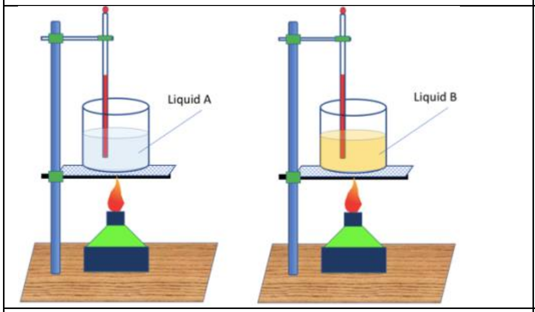
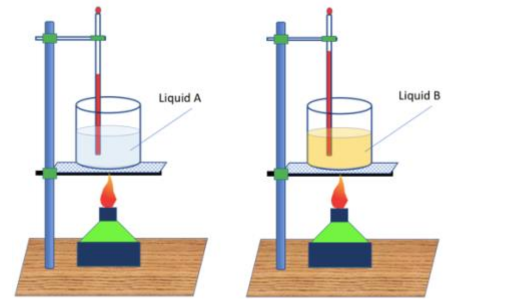
B. Có mối quan hệ nhưng không rõ ràng giữa kích thước và số đốm trên lưng rùa.

C. Không có mối quan hệ nào giữa kích thước và số đốm trên lưng rùa.

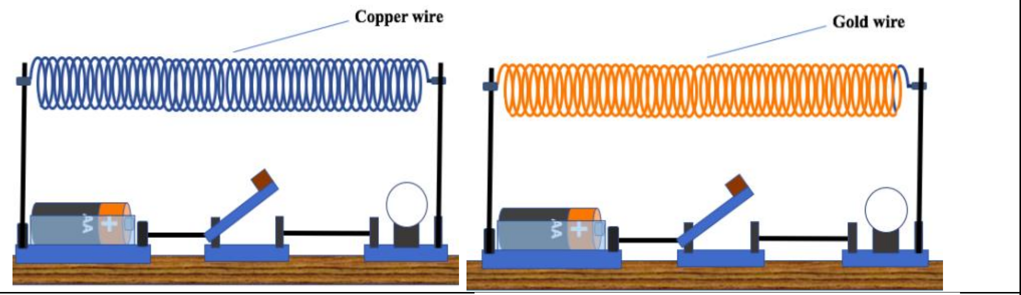
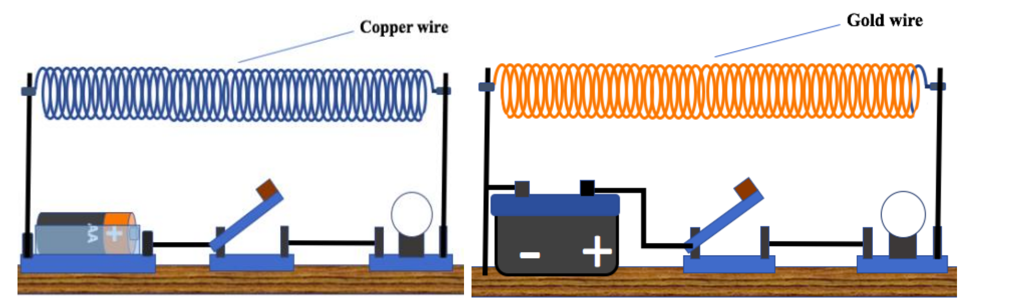
D. Không đủ thông tin để đi đến kết luận.

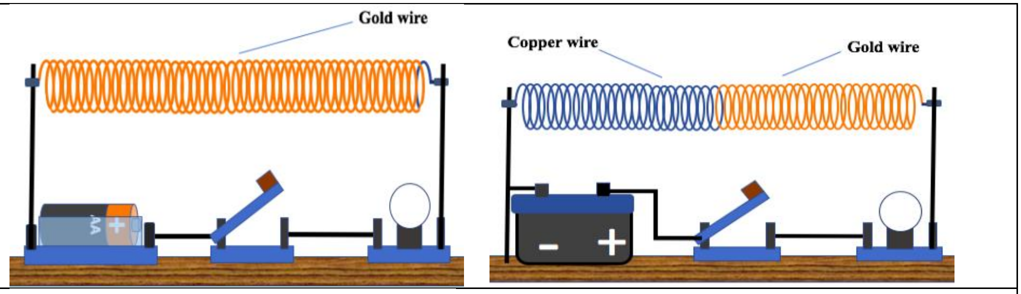
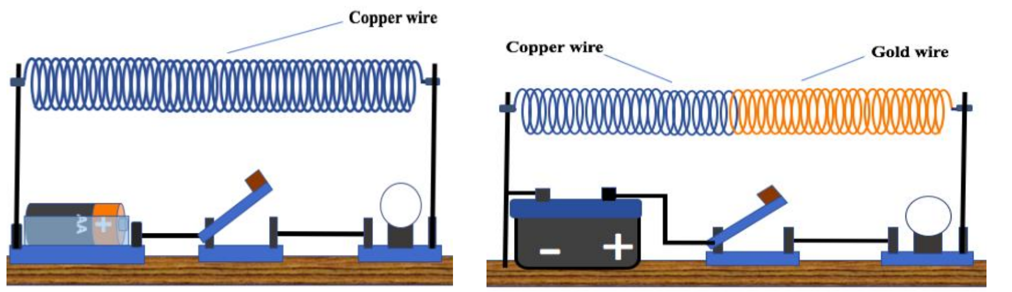
Appendix D. New items developed in the CVSP test

ID03. Nam has two types of liquid: liquid A and liquid B. He wants to determine which one will reach 60°C more quickly in the same initial condition of volume and temperature. The following experiments illustrate four experimental systems in initial conditions. Which one would allow him to test his assumption?

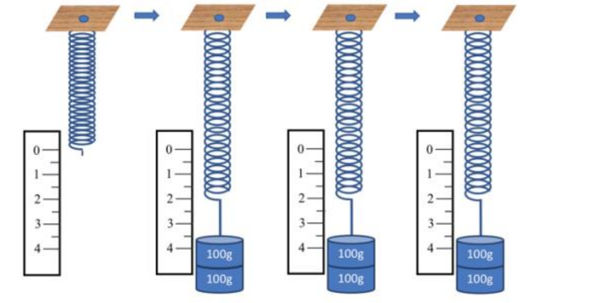
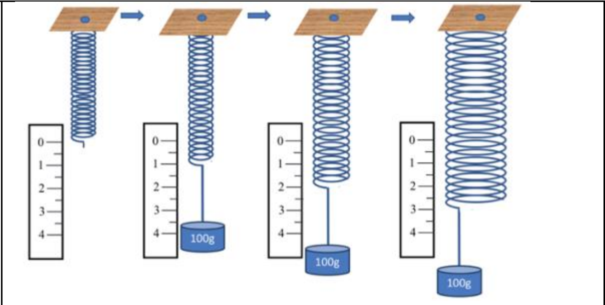
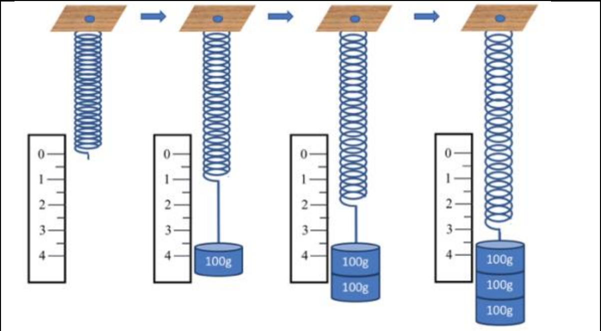
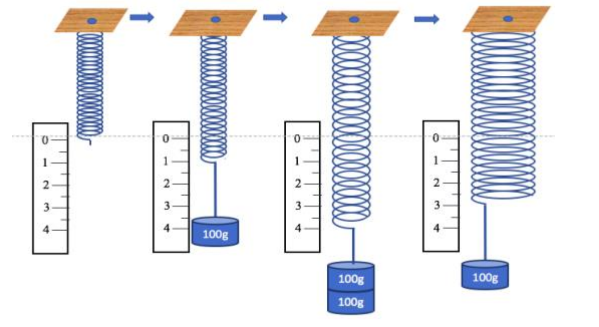
	
A	B
	
C	D

ID05. Mr. Son wants to find out whether the material of a wire has an impact on its resistance. He assumes that the bulb will shine brighter when he uses gold instead of copper to connect it with a battery.
Which the following experimental sets would be best to allow him to test his assumption?

A	
B	

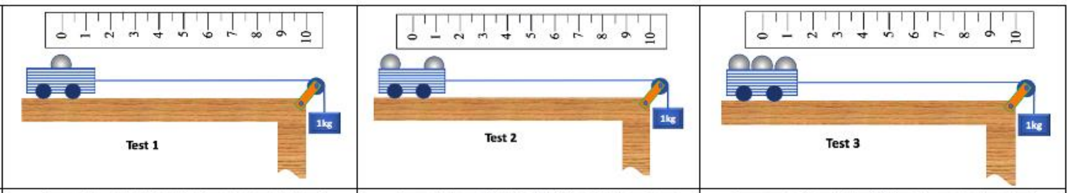
C	
D	

ID07. A group of students has three different springs made of steel, but their natural lengths are equivalent. Students want to determine whether the displacement (stretch) of springs depending on their characteristics. Which set of tests should be conducted?

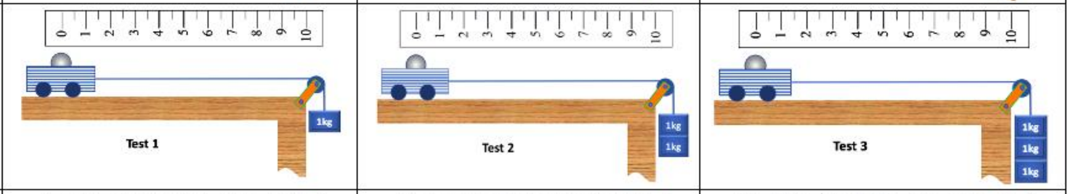
	
A	B
	
C	D

ID08. Mr. Long supposes that if a constant force acts on a cart from rest, the cart’s mass may affect the time that it takes to complete a constant distance. Which set of the following experiments can test his assumption?

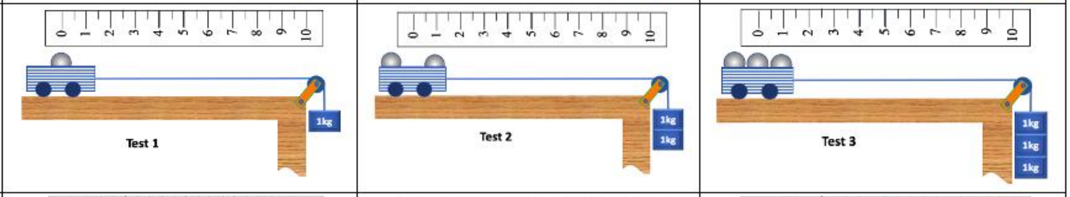
A



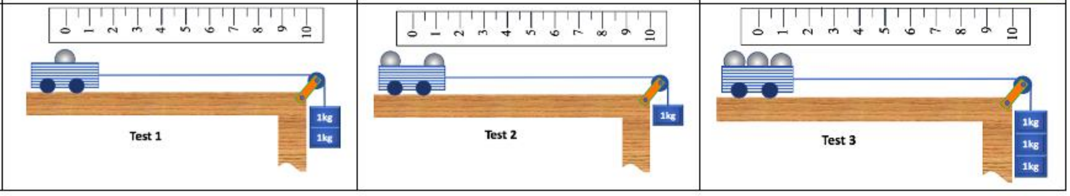
B



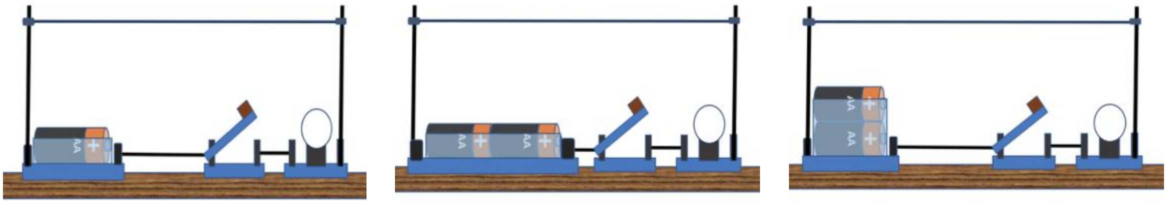
C



D



IN02. A student did the following experiment:

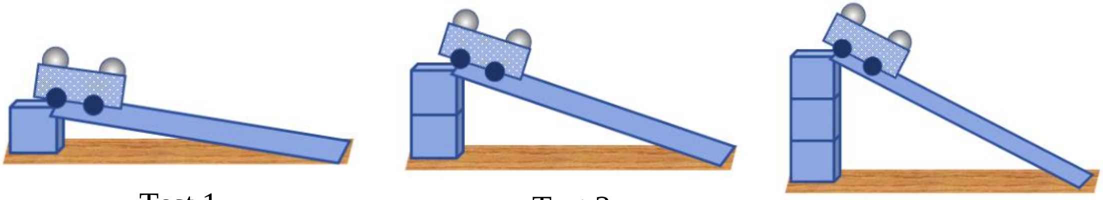


Test 1 Test 2 Test 3

What does her experiment show?

A. The battery and the wire’s material have an impact on the brightness of the bulb.
B. The number of batteries has an impact on the brightness of the bulb.
C. Ways of connecting the batteries impact the brightness of the bulb.
D. The experiment does not allow any valid conclusion.

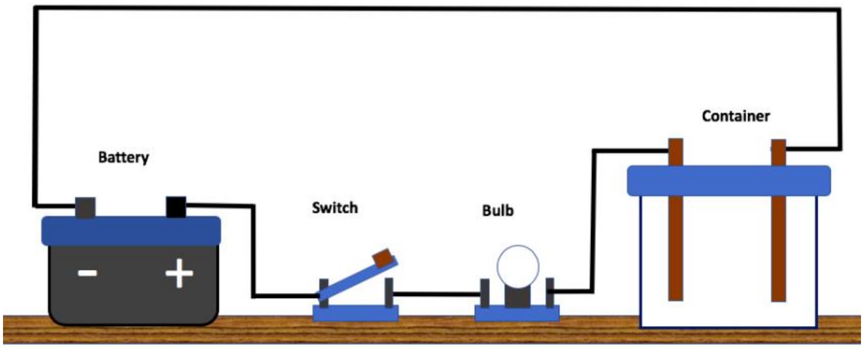
IN04. Minh did the following experiment:



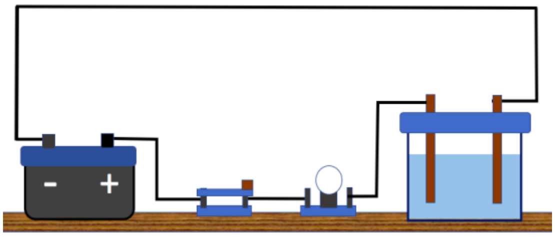
Test 1 Test 2 Test 3

What can he conclude from these experiments?
A. The mass of the cart affects how the cart performs.
B. The height of the ramps affects how the cart performs.
C. The cart’s mass affects how the cart performs, and the height of the ramp affects how the cart performs.
D. It is not possible to reach any valid conclusion.

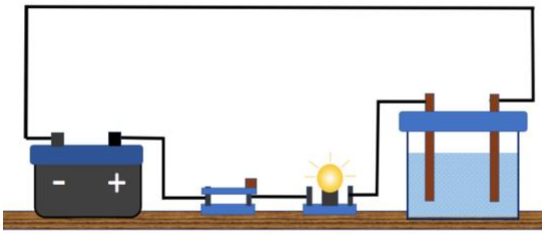
IN08. A group of students set up the following experiments:



They did a series of experiments by filling a container with pure water and then adding salt (NaCl).



Test 1: Still pure water in the container

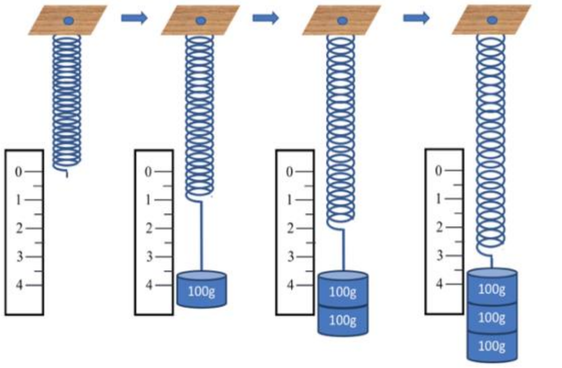


Test 2: Saltwater in the container

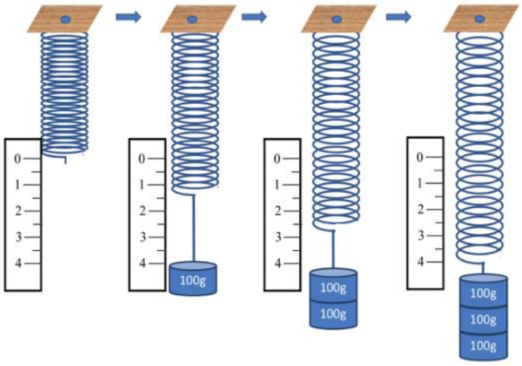
What do their experiments show?

A. Saltwater can carry the current, but pure water cannot carry the current.
B. Pure water carries the current, but saltwater cannot carry the current.
C. The amount of salt in water affects the brightness of the bulb.
D. The experiment does not allow any valid conclusion.

UN05. Miss Thi did the following experiments:



Test 1

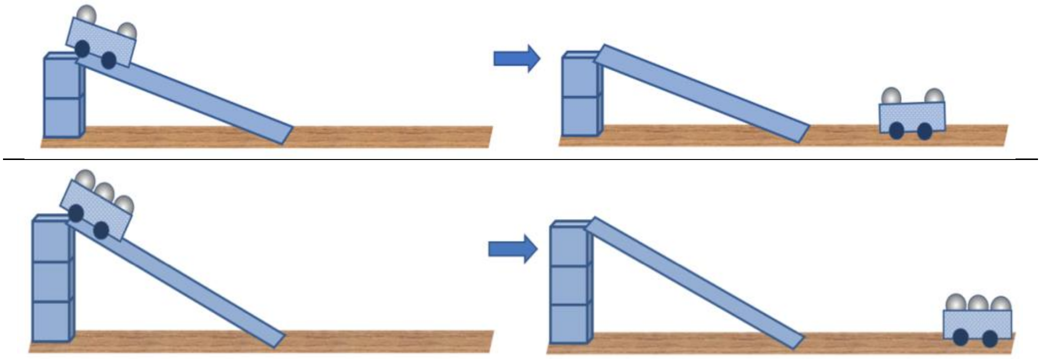


Test 2

What can she find out from the series of experiments?

A. A hanging mass affects the spring’s stretch.
B. The spring’s property affects its stretch.
C. A hanging mass and the property of the spring affect the spring’s stretch.
D. The experiments do not allow us to reach any valid conclusion.

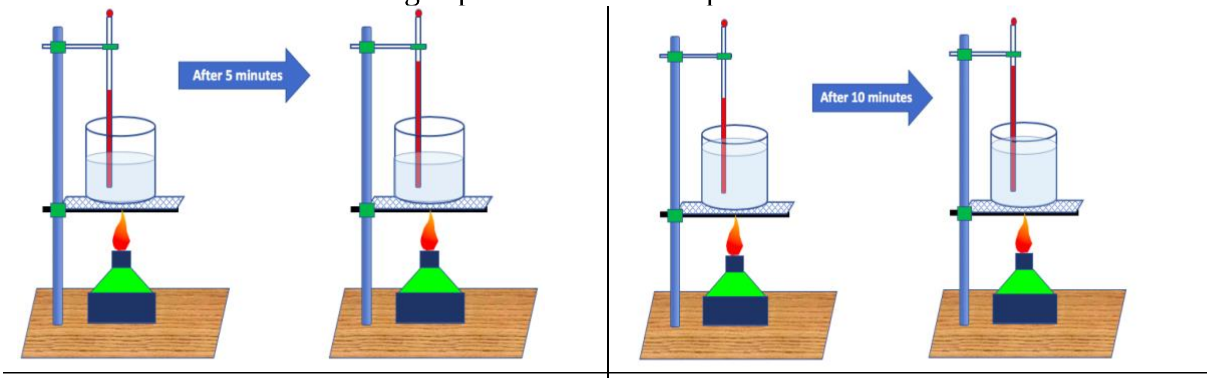
UN07. Peter did the following experiments:



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.

UN08. Mr. An did the following experiments with a liquid:



Experiment 1 | Experiment 2

What can he find out from doing the experiments?

A. Relationship between the amount of heat (energy) that the liquid absorbed and the mass of the liquid.
B. Relationship between the amount of heat (energy) that the liquid absorbed and temperature rise.
C. Relationship between the amount of heat (energy) that the liquid absorbed and the property of the liquid.
D. The experiment does not allow us to deduce any valid conclusion.

Appendix E. The psychometric properties of the IR test

Item	Task	Correct answer (%)	Discrimination	Difficulty	Infit MNSQ
1	FS	83.61	0.51	−0.69	0.91
2	FS	87.33	0.52	−1.04	0.90
3	FS	81.68	0.50	−0.53	0.94
4	FS	90.77	0.40	−1.45	0.96
5	FS	30.72	0.32	2.31	1.10
6	FS	89.26	0.39	−1.26	1.01
7	FA	55.23	0.43	1.04	1.07
8	FA	56.47	0.46	0.98	1.04
9	FA	75.90	0.46	−0.11	0.98
10	FA	74.66	0.51	−0.03	0.97
11	NA	79.75	0.37	−0.38	1.06
12	NA	85.67	0.47	−0.88	0.92
13	NA	83.47	0.42	−0.68	1.00
14	NA	78.93	0.52	−0.32	0.96
15	NA	38.57	0.28	1.88	1.12
16	NS	89.39	0.48	−1.27	0.87
17	NS	71.63	0.52	0.16	0.94
18	NS	68.46	0.46	0.34	1.02
19	NS	58.68	0.55	0.87	0.96
20	NS	54.82	0.50	1.06	1.01

Appendix F. The psychometric properties of the SR test

Item	Tasks	Correct answer (%)	Discrimination	Difficulty	Infit MNSQ
1	CO	58.26	0.43	0.23	1.00
2	CO	71.18	0.41	−0.45	1.01
3	CO	22.59	0.21	1.94	1.02
4	SS	77.55	0.31	−0.76	1.03
5	SS	72.31	0.42	−0.45	0.97
6	SS	69.42	0.37	−0.30	1.02
7	CL	64.05	0.40	−0.04	0.99
8	CL	36.64	0.24	1.20	1.08
9	CL	76.72	0.41	−0.71	0.99
10	AS	68.46	0.34	−0.25	1.02
11	AS	85.67	0.36	−1.34	0.96
12	AS	85.81	0.46	−1.35	0.95
13	PR	65.15	0.42	−0.09	0.99
14	PR	53.31	0.39	0.45	0.98
15	PR	65.70	0.53	−0.12	0.91
16	CR	46.42	0.30	0.76	1.03
17	CR	67.08	0.35	−0.18	1.03
18	CR	31.40	0.15	1.45	1.10

Note. CO: conservation, CL: classification, PR: proportional reasoning, CR: correlational reasoning, SS: Series completion with science-related content, AS: Analogies with science-related content.

Appendix G. The psychometric properties of the SMTSL questionnaire

Item	Subscale	Content	D	Logit location	Infit MNSQ
1	SE	Whether the science content is difficult or easy, I am sure that I can understand it.	0.49	0.33	1.05
2	SE	I am sure that I can do well on science tests.	0.49	0.52	1.06
3	SE	No matter how much effort I put into it, I cannot learn science.(−)	0.36	0.00	1.22
4	SE	When science activities are too difficult, I give up or only do the easy parts.(−)	0.32	0.28	1.31
5	SE	During science activities, I prefer to ask other people for the answer rather than think for myself. (−)	0.26	0.42	1.43
6	AL	When learning new science concepts, I attempt to understand them.	0.53	−0.23	0.90
7	AL	When learning new science concepts, I connect them to my previous experiences	0.55	−0.28	0.90
8	AL	When I do not understand a science concept, I find relevant resources that will help me.	0.61	−0.39	0.88
9	AL	When I make a mistake, I try to find out why.	0.54	−0.58	0.92
10	AL	When new science concepts that I have learned conflict with my previous understanding, I try to understand why	0.59	−0.30	0.90
11	LV	I think that learning science is important because I can use it in my daily life.	0.51	0.06	1.01
12	LV	I think that learning science is important because it stimulates my thinking.	0.63	0.02	0.85
13	LV	In science, I think that it is important to learn to solve problems.	0.45	0.08	1.08
14	LV	In science, I think it is important to participate in inquiry activities.	0.47	−0.10	1.04
15	LV	It is important to have the opportunity to satisfy my own curiosity when learning science.	0.58	−0.26	0.93
16	AG	During a science course, I feel most fulfilled when I attain a good score on a test.	0.46	−0.14	1.10

17	AG	During a science course, I feel most fulfilled when I am able to solve a difficult problem.	0.59	−0.47	0.89
18	AG	During a science course, I feel most fulfilled when the teacher accepts my ideas.	0.57	−0.06	0.93
19	AG	During a science course, I feel most fulfilled when other students accept my ideas.	0.49	0.18	1.00
20	LE	I am willing to participate in this science course because the content is exciting and varied.	0.64	−0.02	0.82
21	LE	I am willing to participate in this science course because the teacher uses a variety of teaching methods.	0.62	0.20	0.89
22	LE	I am willing to participate in this science course because the teacher does not put a lot of pressure on me.	0.42	0.52	1.19
23	LE	I am willing to participate in this science course because it is challenging.	0.66	−0.15	0.84
24	LE	I am willing to participate in this science course because the students are involved in discussions.	0.48	0.36	1.06

Note. SE: self-efficacy, AL: active learning strategies, LV: science learning value, AG: achievement goals, LE: learning environment stimulation; (−): Reverse coded items; D: Discrimination.

RELEVANT PUBLICATIONS

Journal articles

Van Vo, D., & Csapó, B. (*in press*). Measuring inductive reasoning in school contexts: A review of instruments and predictors. *International Journal of Innovation and Learning*.

Van Vo, D., & Csapó, B. (2021). 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*. https://doi.org/10.1080/09500693.2021.1957515

Van Vo, D., & Csapó, B. (2021). 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. (2020). Development of inductive reasoning in students across school grade levels. *Thinking Skills and Creativity*, *37*(2020), 100699. https://doi.org/10.1016/j.tsc.2020.100699

Conference papers

Van Vo, D. (2021). Interactions of Reasoning Abilities and Science Motivation with Parent Involvement Variables in Predicting the STEM Achievement. In: Molnár, Gyöngyvér; Tóth, Edit (ed.): *The answers of education to the challenges of the future: XXI. ONK*. National Educational Science Conference. November 18-20, 2021. Szeged, Hungary: Institute of Education, University of Szeged, pp. 218-218.

Van Vo, D., & Csapó, B. (2021). A Comparison between Technology-Based and Paper-Based Assessment on the Scientific Reasoning Test in Control of Variables Strategy in Physics. In: Molnár, Gyöngyvér; Tóth, Edit (ed.): *The answers of education to the challenges of the future: XXI. ONK*. National Educational Science Conference. November 18-20, 2021. Szeged, Hungary: Institute of Education, University of Szeged, pp. 167-167.

Van Vo, D.(2021). Development of inductive reasoning and its relationship with science performance in Vietnam. Abtract book: *The 14th Training and Practice International Conference on Educational Science*. Kaposvár, Hungary: Kaposvár University Faculty of Pedagogy, pp. 87-87.

Van Vo, D. (2021). Exploring the patterns of inductive reasoning, scientific reasoning and science motivation in Vietnamese students across grade levels. Abstract book: *In EARLI*

2021: Education and Citizenship: Learning and Instruction and the Shaping of Futures. Online, pp. 230-230.

Van Vo, D. (2021). Assessing control of variables strategy in physics of high school students. Abstract book: *In EARLI 2021: Education and Citizenship: Learning and Instruction and the Shaping of Futures*. Online, pp. 305-305.

Van Vo, D. (2020). Motivation toward science learning and inductive reasoning ability of secondary students in Vietnam: developmental changes and relationships. Abstract book: *VI. Ipszilon Konferencia*. Budapest, Hungary: ELTE PPK, pp. 48-49.

Van Vo, D. (2020). Assessment of motivation toward science learning and scientific reasoning of students in Vietnam: developments and relationships. Abtract book: *The 13th Training and Practice International Conference on Educational Science*. Kaposvár, Hungary: Kaposvár University Faculty of Pedagogy, pp. 111-111.

Van Vo, D. (2019). Technology-based assessment of reasoning skills: A literature review. In: Varga, Aranka; Andl, Helga; Molnár-Kovács, Zsófia (eds.) Neveléstudomány – Horizontok és dialógusok. Absztraktkötet. : *XIX. ONK*. Pécs, 2019. november 7-9. Pécs, Hungary : Pécsi Tudományegyetem Bölcsészettudományi Kar Neveléstudományi Intézet, pp. 383-383.

Van Vo, D. (2019). *Technology-based assessment in the educational setting.* In: Molnár, Edit Katalin; Dancs, Katinka (eds.) *CEA 2019: 17th Conference on Educational Assessment* . Programme and Abstracts. Szeged, Hungary: Szegedi Tudományegyetem, pp. 98-98.

DECLARATION

I hereby certify that the content of this dissertation is my original work of production. This dissertation has not been submitted for any other degree previously or at any other educational institution.