Development of Tools and Assessment of Ethiopian Students' Pattern of Scientific Reasoning at the Upper Primary Level in the Hadiyya Zone

Tsedeke Abate

Addis Ababa University

Department of Science and Mathematics Education

December, 2020

Development of Tools and Assessment of Ethiopian Students' Pattern of Scientific Reasoning at the Upper Primary Level in the Hadiyya Zone

Tsedeke Abate

A PhD Dissertation Submitted to the Department of Science and Mathematics Education

Presented in Fulfillment of the Requirement for the degree of Doctor of Philosophy in Physics Education

Addis Ababa University

Department of Science and Mathematics Education

December, 2020

Addis Ababa University

College of Education and Behavioral Studies

This is to certify that the dissertation prepared by **Tsedeke Abate**, entitled: Development of Tools and Assessment of Ethiopian Students' Pattern of Scientific Reasoning at the Upper Primary Level in the Hadiyya Zone and submitted in fulfillment of the requirements for the Degree of Doctor of Philosophy in Physics Education complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Assessment Committee

____________________ External Examiner	____________________ Signature	________________ Date
____________________ Internal Examiner	____________________ Signature	________________ Date
___ Professor Carl Angell Main Supervisor	____________________ Signature	________________ Date
_________Dr. Kassa Michael ______ Co-Supervisor	____________________ Signature	________________ Date
____________________________________ Department Chair/Coordinator	____________________ Signature	________________ Date

Abstract

Development of Tools and Assessment of Ethiopian Students' Pattern of Scientific Reasoning at the Upper Primary Level in the Hadiyya Zone

Tsedeke Abate

Addis Ababa University, 2021

Studies indicate that the failure of science education to meet the needs of the 21st century is to some extent due to the inability to incorporate scientific reasoning and higher order assessments in the school instruction. Though the outcomes of education seek higher-order thinking abilities there is a lack of high ability assessments in low-income nations. The purpose of this mixed methods research is to explore students' way of reasoning and develop and validate the Scientific Reasoning Progress Tool (SRPT) which measures students' reasoning abilities. The phenomenographic study was used to explore grade eight students' patterns of reasoning, and to validate an assessment tool the descriptive survey method was used. The qualitative analysis of students' interviews elicited three major themes of student views towards different levels of scientific reasoning: naïve, mixed, and scientific along with the underlying ways of reasoning. For the quantitative part of the study, 40 items were developed, pilot tested, and administered for 242 grade eight students at six schools. Rasch Analysis was used to develop and validate the items with the help of Winsteps. It was found that the SRPT is offered as a valid and reliable instrument that is a construct-driven, domain-specific, and a higher-order thinking assessment that can measure students' reasoning progress. However, it was found that the reasoning ability of grade 8 students' is limited to lower level reasoning and they faced difficulty in attaining higher-level reasoning like

generating evidence and drawing conclusions. This study has produced a valid and reliable assessment tool to measure students' scientific reasoning abilities and a theoretical model that could be used to develop assessment tools. The framework and design of this research may be adapted to develop additional instruments to investigate the learning progression of students' scientific reasoning in middle and high school classes in Ethiopian and other low-income countries.

Dedicated to

Inner Soul

Acknowledgments

I would like to thank my former research supervisor, Associate Professor Per Morton Kind for his loving-kindness and compassion he showed and helping me to grasp the basics and discover the research area. I would like to thank my research supervisor, Professor Carl Angell for his useful comments, and patience throughout my candidature. I appreciate him for reading patiently all the materials I sent him with constructive comments. I would like to thank my internal research supervisor, Associate Professor Kassa Michael for his useful comments, advice, and patience throughout my candidature. I appreciate him for the tremendous patience he showed upon me even when I was struggling to continue. I would like to thank Dr. Mekbib Alemu for his comments and advice from the inception of the program to the end of this dissertation work. I'd would like to express my appreciation to Dr. Getachew Tarekegn for his contributions, comments, and suggestions. I would like to thank Adisu Daniel for his comments. I would like to thank my colleagues Dr. Hassen Worku, and Dr. Deresse Terfe for their contribution by sharing their ideas. I would like to thank Teshome Forsido, Gedeno Mollo, Ageze Abza, Adal Jekamo, Ertembo Mishore, and Amaregn Kassaye for their comments. I would also like to thank Fisseha Fentaye and his family for his advices and financial support. Finally, I would like to thank my wife Endalu Kelklachew, and my kids: Soteria and Shalom for their tolerance and support when I was occupied with my study.

Publications, Conferences, and Proceedings from this Ph. D. dissertation

Publications

Abate, T., Michael, K, & Angell, C. (2020). Assessment of Scientific Reasoning: Development and Validation of Scientific Reasoning Assessment Tool. *EURASIA Journal of Mathematics, Science and Technology Education*, 16(12), em1927. *https://doi.org/10.29333/ejmste/9353*

Abate, T., Michael, K., & Angell, C. (2021). Upper Primary Students' Views Vis-à-Vis Scientific Reasoning Progress Levels in Physics. *Eurasia Journal of Mathematics, Science and Technology Education*, *17*(5), em1958. https://doi.org/10.29333/ejmste/10834

Definitions of terms

1. Classical test theory (CTT) is a test theory which provides psychometrics about a test such as reliability of a test.

2. Constructs are the variables to be studied in a specific study i.e. the construct for this study was scientific reasoning.

3. Construct map is a well thought out ordering of qualitatively different levels of students' scientific reasoning ability.

4. Item response theory (IRT) is a test theory which provides performance of an individual and how it relates to individual items.

5. Assessment tool is set of items designed to measure scientific reasoning in this study.

6. Scientific reasoning progress is a way and pattern one develops his ability to reason from low (naïve) to high reasoning level.

7. Scientific styles of reasoning are six styles in which they were developed considering and analyzing historical development of science knowledge in the course human reasoning development.

8. Proficiency level is a progress of students' understanding a concept from low level of understanding to high level of understanding.

9. Scientific reasoning is a process of thinking which involves generating a claim, providing explanation with valid evidences either to accept or reject a claim in order arrives at valid conclusion.

Contents

List of Tables

List of Figures

List of Acronyms

AERA	American Educational Research Association
APA	American Psychological Association
BAS	BEAR Assessment System
BEAR	Berkeley Evaluation and Assessment Research
CPD	Continuous Professional Development
CTT	Classical test theory
EGRA	Early Grade Reading Assessment
HDP	Higher Diploma Programs
ICDR	Institute of Curriculum Development and Research
IRT	Item Response Theory
LCTSR	Lawson's Classroom Test of Scientific Reasoning
LP	Learning Progression
NAEP	National Assessment of Educational Progress
NEAEA	National educational assessment and examination agency
NOE	National Organization for Examination
PISA	Program for International Student Assessment
SIP	School Improvement programs
SRPT	Scientific Reasoning Progress Tool
SSR	Styles of Scientific Reasoning
TDP	Teachers Development Program
TIMSS	Trends in International Mathematics and Science Study
USAID	Untited States Agency for International Development

Chapter 1 Introduction

1.1 Background of the study

In recent years science education has been recognized as crucial to transform and to bring overall development and prosperity to nations. For example, sub-Saharan and Southern American countries place emphasis on science education in the hope that its outcomes will help them to reduce poverty and to improve the poor life standards (MoE & JICA, 2016), whereas other nations hope that science education will enable them to alleviate their multidimensional political crisis, to bring about economic development (Han, 2013; Stacey et al., 2018; Turiman, Omar, Daud, & Osman, 2012), to improve their citizens' standard of life, and to solve their socio-scientific and environmental problems (Zhou et al., 2016). To overcome these challenges, science education needs to develop students' scientific reasoning abilities, which include the ability to solve problems, to evaluate information, to work collaboratively, to apply new technology, to think critically, and to reason about nature scientifically (Dole, Bloom, & Kowalske, 2016; Lamb, Jackson, & Rumberger, 2015; Nagaoka, Farrington, Ehrlich, & Heath, 2015). Around the globe, science education is considered as crucial to equip the next generation with the necessary scientific literacy skills that enable them to be the future professional in science and technology fields as well as to be competent, critical, and informed decision-makers and consumers of the product of science and technology.

Ethiopia is a low-income nation with a desire to become a middle-income country by 2025. There is a huge demand for creative and innovative solutions to provide quality services and to sustain the existing development. These creative and innovative approaches require higher-order thinking and reasoning skills. As a result, Ethiopia requires a scientifically literate citizen with relevant knowledge and understanding about scientific concepts and processes. In line with this, Ethiopian science education aims to train future professionals in science and technology as well as the next generations who are informed decision-makers about socio-scientific issues and also skilled to judge and question information and products of science and technology.

Recently, scientific reasoning has gained attention in STEM (Science, Technology, Engineering, and Mathematics) education due to the potential of reasoning skills to develop students understanding of the nature of science and its practices (Kind & Osborne, 2017; Opitz, Heene, & Fischer, 2017). Studies also indicate that scientific reasoning promotes students' problem-solving skills in a real-life situation (Han, 2013). In addition to this, scientific reasoning affects students' achievement and success in science education (Bao et al., 2009; Osborne, 2013). In science instruction, nowadays, there's is a shift towards teaching strategies that promote students' scientific reasoning and critical thinking. As the need for teaching scientific reasoning skills is recognized, there is also an increasing demand for such assessments. In brief, the emphasis given to scientific reasoning ability in science instruction should be paralleled by its assessments too. Students' scientific reasoning skills can be promoted only when the assessment mirrors the instruction and assesses students' level of reasoning. Moreover, such assessments should be employed continuously to gain valid information about students' levels and abilities of scientific reasoning (NRC, 2000).

A review of science education studies reveals that relatively few studies have been conducted to assess students' scientific reasoning, although the importance of scientific literacy has been widely discussed in science education and psychological research (Opitz et al., 2017). Most of the existing assessments were focused on dimensions of scientific reasoning such as *problem identification, questioning, hypothesis generation, evidence generation, evidence evaluation,* and *drawing conclusions*. In addition, some of these tests assumed scientific reasoning as a single general ability whereas others considered scientific reasoning as a consortium of several independent skills. Although some studies focus on scientific reasoning in its general sense (Lawson, 2004), recent trends depict a shift towards domain-specificity of scientific reasoning skills. Most of these assessment tools had reported certain psychometric properties of their tests. The most frequently reported ones are reliability, content validity, construct validity, and criterion validity.

Science educators have only recently begun to raise concerns about the scientific knowledge that scientific reasoning relies on. Content, procedural, and epistemic knowledge are necessary for scientific reasoning (Kind & Osborne, 2017; Osborne, Rafanelli, & Kind, 2018). There is, however, a lack of valid and reliable test instruments that can uncover students' epistemic knowledge in physics.

Several international and national scientific reasoning tests have been developed to assess students' level of reasoning (Opitz et al., 2017). Among them, Lawson's Classroom Test of Scientific Reasoning (CTSR), PISA, TIMSS, PIRLS, and NAEP are the most prominent ones. The CTSR is one of the assessments widely used by researchers in science education (Bao et al., 2009; Lawson, 2004). The test measures various aspects of scientific and mathematical reasoning. That indicates that the test measures the general characteristics of scientific and mathematical reasoning. The Trends in International Mathematics and Science Study (TIMSS) has been engaged in developing items and assessing students' abilities based on the three cognitive aspects: knowing, applying, and reasoning. The Programme for International Student Assessment (PISA) is also engaged in developing items to measure scientific literacy. The National Assessment of Educational Progress (NAEP) is also one of the best known American assessment framework projects in developing assessment which includes procedural and epistemic knowledge embedded in science performance (Kind, 2013).

However, Osborne (2013) argued that domain-general scientific reasoning tests neglect contextual considerations such as children's ways of learning in relation to different forms of scientific knowledge, and conceptualizations of a construct to be assessed. As a result, Anderman, Sinatra, and Gray (2012) argued that such assessments are limited to accurately evaluating the thinking skills and understanding of students by taking into account particular learning circumstances relevant to a specific topic in detail. Besides, various studies considered science learning as domain-specific (science-as-practice) not as domain-general (science-as-logic), and the developed scientific knowledge as progressive (Gelman & Noles, 2011; Lehrer & Schauble, 2015; Osborne & Wittrock, 1985). The science and mathematics assessments of PISA, TIMSS, and NAEP, which are developed for the various grade level and administered around the globe, considered scientific reasoning as one theme of the science test. However, these tests are criticized since they give much emphasis to the domain-general aspect of scientific reasoning (for example TIMSS), neglect knowledge and reasoning dimensions of science (for example NAEP), fail to explain the link between knowledge dimensions and scientific reasoning (for example PISA) (Kind, 2013). Thinking that, these assessments were developed under the setting of developed countries and used to testify about the effectiveness of a country's educational system and its education quality, questions have been raised whether the scientific reasoning items were context-dependent and culturally situated for low-income nations like Ethiopia. Furthermore, the objectives of science education vary from countries to countries based on its demand and level of economy (Kambeyo, 2017).

Another limitation of previous assessments of scientific reasoning was that most of the tests not only focused on skills such as *variable relationship, hypothesis generation, constructing experiment, evidence evaluation,* and *drawing of conclusion* but also neglected different forms of knowledge (Kind & Osborne, 2017). Studies have shown that children develop domain-specific theories that, in turn, lead to domain-specific reasoning from childhood through adulthood. Besides, it has been found that pre-school children and elementary school students participate in reasoning and higher-level tasks using domain-specific knowledge (Gelman & Noles, 2011). These studies also indicate that human cognition is closely related to domain-specific knowledge, such as the reasoning ability of students, and has a progressive developmental process starting at an early age. Therefore, because such a stage is attributed to grasping the basics of physics knowledge, it is important to consider the domain-specific scientific reasoning capacity of primary and middle school students. Besides, literature in science education has introduced a fresh and detailed conceptualization of scientific reasoning (see (Kind & Osborne, 2017; OECD, 2013; Osborne et al., 2018). Kind and Osborne (2017) approach scientific reasoning from the perspective of the cognitive history of science and raise issues concerning the scientific knowledge on which scientific reasoning depends. Scientific reasoning requires content, procedural and epistemic knowledge. This failure of previous reasoning tests to consider different forms of knowledge, may originate from their failure to provide a “clear and coherent conceptual model for the sciences” (Osborne et al., 2018, p. 16). Such assessments measure students' reasoning based on the reflections of the thinking process instead of understanding scientific reasoning directly from the science culture where the reasoning has originated (Kind, 2013).

Although previous tests have reported some psychometric properties, there is limited research that ha comprehensively reported the method of analysis of the results. Many assessments use classical test theory (CTT) to analyze the responses of students which only provides statistical meaning. CTT has an assumption that all test items make the same contribution to the overall score independently of their individual item difficulties and students' abilities. In education, it is particularly important that tests support good teaching and learning, and that tests assess what educators think is important (Kind, 2013). Assessors, therefore, strive for “construct-driven” and “evidence-based” assessments that support learning progression. Modern test theories allow construct-driven assessment procedures that help to develop progressive levels of a construct (Kind, 2013; Wilson, 2005).

African countries favor assessment of higher-order thinking ability such as synthesis, analysis, and evaluation, however, the school made science assessments are heavily dominated by test items that measure lower cognitive abilities and conceptual knowledge (Arega, 2014; Verspoor, 2008). According to Arega (2014) in the Ethiopian context, different agencies have tried to prepare assessment guidelines, techniques, and resource materials that focus on objective and content. Even though the Ethiopian curriculum framework encourages students' dialogue, argumentation, and higher-order knowledge, this was not observed in both national and classroom examination of science subjects (Teshome, 2017). Previous tools are mainly aimed either to compare the score with some established criteria or to check the attainment of stated objectives. They are not based on conceptualizing a construct (construct-based) and considering the evidence-based argument for they do not consider the analysis of theoretical frameworks, analysis of students' alternative conceptions, and context-based students' learning before item creation. This indicates that the existing assessment in the Ethiopian context is item-driven assessment instead of considering construct-driven assessments. In summary, to promote the teaching of scientific reasoning and to make it a common feature of science classrooms, science education, in the world in general and in Ethiopia in particular, is faced with a challenge of defining progression in the competency of scientific reasoning and providing items that assess student reasoning and thinking in science subjects.

Therefore, there is a paucity of scientific reasoning test instruments that can uncover students' higher-order thinking progressively in elementary school physics subjects. Particularly, a scientific reasoning test, in which the construct 'scientific reasoning' is conceptualized and its items are context-based, is needed in the Ethiopian context. Thus, this study attempted to conceptualize scientific reasoning and discussed the development and validation of an instrument that assesses students' physics scientific reasoning - namely Scientific Reasoning Progress Tool (SRPT hereafter).

Besides, this study investigates the progression of grade 8 students' reasoning using the SRPT. The tool developed here differs from previous such tools in many regards. One, it is context-based in which it considers the contextual factors by participating students and middle school physics teachers in developing and validating the progress levels and the tool as well. Two, SRPT used the *Styles of Scientific Reasoning* framework to identify and demarcate the scientific reasoning construct. The proficiency level within the components of the *Styles of Scientific Reasoning* was also addressed. Three, we followed the Wilson assessment cycle and Senocak's (2009) framework for the development of SRPTs' items as well as to establish the scale. In short, the test development and its validation employed the modern test theory, commonly known as Item Response Theory (IRT) instead of dealing specifically with the classical test theory (CTT).

1.2 Statement of the problem

This study aspired to address two long-standing educational problems. One is related to measuring scientific reasoning skills at different levels. This has been a problem since science commenced as a school subject and science educators divided the science curriculum into the scientific method and science knowledge (Dunbar & Klahr, 2012; Ford & Forman, 2006; Kind & Osborne, 2017). Students, it was claimed, should not just learn concepts and theories scientists used to explain the world but also the method (Dunbar & Klahr, 2012). The two competing models, science as a body of knowledge and science as a practice, dominated in science curriculum since then (Ford & Forman, 2006), where the two models have weaknesses in that they do not consider the impact of cultural value and influence. Since the second half of the 20th century, however, the understanding of scientific reasoning has changed substantially, with a major impact on science education (Dunbar & Klahr, 2012; Kind, 2013). Science, it is claimed, is a social activity based on argumentation and discourse between scientists to decide what concepts and theories better fit the natural world. Scientific reasoning, accordingly, is a mixture of following standardized procedures (as a process), using socially established epistemic criteria, and building on established scientific knowledge (as a product); all in a context of debate and discourse (Driver, Newton, & Osborne, 2000; Ford, 2015; OECD, 2013; Osborne, 2013). This new development makes it necessary to rethink how we teach scientific reasoning in science education, and also how it is assessed. Following the change in the focus and approach towards science education, countries started to incorporate scientific reasoning and argumentation in their standards and devise the assessment of such abilities by developing standardized tools (OECD, 2012). There is a global shift towards students learning skills and literacy. One of the skills focused on global and sub-Saharan African countries is scientific reasoning. However, the effort is yet in progress globally and very limited in low-income nations like sub-Saharan African countries (Altinok, Angrist, & Patrinos, 2018; Bloem, 2013; Verspoor, 2008). Consequently, there is growing interest globally to assess the students' pattern of reasoning and their reasoning ability levels of low-income nations. But the work done so far is either still limited or not addressed at all in primary schools of sub-Saharan African countries' (Cresswell, Schwantner, & Waters, 2015; OECD, 2015). Even though most sub-Saharan African countries included skills of scientific reasoning in their curriculum, the students' ability of such skills was not explored and assessed properly (Joshi & Verspoor, 2013; Verspoor, 2008).

The second problem is related to the assessment tool in low-income nations. The school assessment system of low-income nations (especially sub-Saharan African countries) is dominated by lower ability levels. In many sub-Saharan African countries, current classroom assessments are limited in measuring students' low-level understanding such as recall, memorization of facts, and ability to use algorithms (Joshi & Verspoor, 2013; Verspoor, 2008). One may argue that there are international assessment tools by projects such as TIMSS, NAEP, and PISA to solve such problems and improve the educational system. However, studies reveal that the students of low-income nations who participate in such international tests lag behind high-income nations in educational achievement. International studies, such as TIMSS, PIRLS, and PISA show an achievement gap of about two standard deviations between the international mean value of these tests and the score in a typical low-income nation (Martin & Jamieson-Proctor, 2019; OECD, 2016; Sabanathan, Wills, & Gladstone, 2015). It means that the average student in low-income nations has four to five year learning lag behind a similar student in the high-income nations (Sabanathan, Wills, & Gladstone, 2015). International tests are typically developed in high-income nations, which means, their suitability for low-income nation students are limited (Greaney & Kellaghan, 2008). The problem is partly that international tests have few items matching the ability level of the average students' low-income nation (since these students are two standard deviations below the international mean and most items are made to match the mean), and partly that low-income nation students live in a very different culture than those in high-income countries (Greaney & Kellaghan, 2008). It is difficult to develop a common assessment procedure that suits all countries for there are curricular variations, achievement gaps, and weight given for a particular skill in terms of various interests of the country, in particular, to low income nations. For this reason, international assessors are increasingly focusing on this problem and have pointed to a demand to develop new assessments better suited for low-income nations (Hanushek & Woessmann, 2015).

The research project Transforming Pedagogy of STEM Subjects (TPSS) has been working the classroom dialogue with scaffolding activities. TPSS is a research project that investigates ways of improving the pedagogy of STEM teaching in Ethiopia. The research study has found that Ethiopian students have low conceptual understanding and reasoning ability when they are compared with their counterparts of high and middle-income nations (Alemu, Kind, Tadesse, Atnafu, & Michael, 2017). However, the items for the study (TPSS) were selected from the international assessment tool, TIMSS. The reason for the students' low conceptual and reasoning ability might be related due to the items were not developed by considering students' way of reasoning and level of reasoning in the Ethiopian context. This indicates that there is a need to develop items for assessing students' reasoning ability by considering the context of students. In addition to this, the current assessments in Ethiopian and sub-Saharan African countries do not meet the need of the 21st century. It is, therefore, required to investigate students' pattern of scientific reasoning in physics and develop an assessment tool to measure their reasoning levels.

1.3 Objectives of the study

The purpose of this study was to develop a scientific reasoning assessment tool and to explore grade 8 students' patterns of reasoning in physics. The test was developed using a construct map to identify the different levels of reasoning. The construct map, here, is a theoretical representation of students' scientific reasoning ability from a low level to a higher level of reasoning. This construct map was developed by the data collected from the interview of students, teachers' Focus Group Discussion (FGD), and relevant literature. The test items of SRPT represent various levels of reasoning across the construct as demonstrated in the construct map. Moreover, significant efforts were gone to make each item contextual and culturally situated with the help of experts review.

In context, the general objective of this study was to develop an assessment tool for measuring scientific reasoning and to explore students' ways and patterns of reasoning at the upper primary school of Hadiyya Zone. In line with this, the specific objectives of this study are:

- To identify scientific reasoning skills that are addressed by the construct map for the development of scientific reasoning progress tool (SRPT) test items
- To develop test items of Scientific Reasoning Progress Tool (SRPT) to measure grade eight students' scientific reasoning progress;
- To establish content and construct validity of the scientific reasoning progress tool items using a panel of experts;
- To determine the effectiveness of the scientific reasoning progress tool items in measuring grade eight students' scientific reasoning in physics.
- To validate the Scientific Reasoning Progress tool (SRPT) that measure Grade 8 students' scientific reasoning progress.

1.4 Research questions

The purpose of this mixed-methods study was to develop the SRPT items for measuring grade eight students' scientific reasoning skills in physics. The development of these items included a process of content and construct validation in which the researcher sought to determine how well the items measure scientific reasoning. The first study component consisted of students' interviews and FGD with primary school physics teachers to explore students' views in relation to four levels of scientific reasoning progress (*factual knowledge, explanation of factual knowledge, evidence-based reasoning, and drawing* *conclusion*). This yielded a construct map and this, in turn, guided the development of SRPT. The content validity of the SRPT was assured by the teachers and also by a panel of experts. The qualitative data was used to provide validity evidence that suggests the SPRT items are representing the levels of scientific reasoning as it is constructed in the construct map. The second component of this study involved the construct validation by administration and analysis of newly developed items for measuring grade 8 students' scientific reasoning progress.

To sum up, the following research questions guided in addressing the objectives of this study:

1. Which scientific reasoning skills are addressed by the construct map developed for the Scientific Reasoning Progress Tool (SRPT) test items?
2. How are the content and construct validity of the Scientific Reasoning Progress Tool (SRPT) items qualitatively established?
3. How effective are the Scientific Reasoning Progress Tool (SRPT) test items in measuring grade eight students' scientific reasoning in physics?
4. What is the trend of Grade 8 students' progression among the different reasoning subskills measured by the Scientific Reasoning Progress Tool (SRPT) ?

1.5 Significance of the study

There is a common saying that "if we want to change teaching, then we have to change the assessment". To measure students' overall developmental progress there is a need to have high-quality instruments that will measure students' ability to critique the quality of scientific evidence to better support students learning progress. For policymakers and curriculum experts, there has to be empirical evidence about students' progress to take appropriate measures. Curriculum experts and teachers need to know the level of learner to for better improvement.

There is a strong link between assessment and instruction. Assessment tools are required to ensure learning progress by providing proper scaffolding based on identified misconceptions (NRC, 2000). Such tools provide useful information about students' competence, alternative conceptions, and ways to guide students toward an understanding of a concept and develop different levels of reasoning abilities (Griffin & Case, 1997; NRC, 2001). The assessment tool needs to provide teachers to choose proper instructional strategies to support students learning progress and assure the attainment of expected learning performances. It is vital to develop instrument tools that can measure students' reasoning ability so that teachers identify students' ability and possible conceptions to develop instruction that can benefit students to promote their learning. It also gives evidence of students learning for decision making bodies by showing what students can do and cannot do. Students can also benefit by obtaining better support results from instruction that is planned based on the evidence of their real-time performance.

In this regard, the current study in the students' reasoning progresses along with the assessment tool is expected to have importance for science educators, researchers, science education, teachers, students, and educational stakeholders. It will inform teachers about students' ways of reasoning in physics, students' possible alternative conceptions, and provide a way how students' reasoning develops and progresses. The assessment tool will provide teachers a way to develop quality, valid, and higher-order reasoning assessments. It also provides a way of how to revise and validate assessment tools. For the researchers, it provides empirical evidence to conduct further study of students' reasoning ability. The framework and design of this research may be adapted to develop additional instruments to investigate the learning progression of students' scientific reasoning in middle and high school classes in Ethiopian and other African countries. The findings of this study will also give insights for curriculum designers to emphasize inquiry-based education with explicit instruction of the nature of science (NOS) as an instructional approach in science learning to enhance students' reasoning ability. Furthermore, the construct map can also be used for exam and table of specifications developers for grade eight students as empirical evidence about students' reasoning progress.

1.6 Delimitation of the Study

The purpose of this study was to explore students' reasoning, to develop a construct map that represents students' reasoning ability, and to develop and validate an assessment tool for measuring scientific reasoning. Since the nature of the qualitative study requires in-depth information from limited samples, the study was delimited to only Hadiya Zone six primary schools. In this study, about 445 Grade 8 students, and six middle school teachers participated. Using a sequential exploratory mixed research design, students' reasoning patterns and tools in physics were developed. The tool that aimed to represent students' reasoning progress was developed by conceptualizing scientific reasoning in terms of students' progress level and three styles of scientific reasoning: model-based reasoning, data-based reasoning and mathematical reasoning for these are the dominant styles of reasoning in primary school physics (Hacking, 2002).

1.7 Overview of the dissertation

Chapter one deals with the introduction of the problem addressed in this study. It also stated the objectives of the study, and research questions to be answered in this study. Chapter 2 will discuss the relevant review of literature related to assessments in education, assessments of scientific reasoning, and frameworks of scientific reasoning. It will also present the conceptualization of scientific reasoning and conceptual framework that guides this study.

Chapter 3 will present the methods used in this study. It will provide the research perspectives, research design, and sampling techniques. The chapter will also present qualitative and quantitative data gathering and analysis tools and techniques.

Chapter 4 will discuss qualitative analysis. It will present an analysis of qualitative data to answer the first two research questions stated in chapter one.

Chapter 5 will present a quantitative analysis. It provides the quantitative analysis using Rasch analysis to answer the three research questions stated in chapter 1.

Finally, Chapter 6 will provide a discussion of both qualitative and quantitative results vis-à-vis of literature. It will also present important conclusions and recommendations for further study.

Chapter 2 Review of Related Literature

These days, there is a demand for conceptualization and a need for the assessments of scientific reasoning. This chapter will present the construct scientific reasoning using relevant literature and scientific reasoning will be conceptualized. The construct scientific reasoning will be presented in terms of learning progression, the three forms of scientific knowledge, and the styles of reasoning. Following this, frameworks for scientific reasoning will be discussed and a conceptual framework that guides this study will be established, and this will, in turn, guides the development of scientific reasoning tool (SRPT). Moreover, the currently available assessments in global, sub-Saharan African, and Ethiopian contexts in terms of assessments' consideration of scientific reasoning will be discussed. Finally, assessment procedures and types of tests will be discussed.

2.1 Introduction

One of the core skills of the 21st century and goals of science education is considered to be developing students' scientific reasoning (Fischer et al., 2014) along with proper views about scientific knowledge (Billingsley & Fraser, 2018; Fernandes, Rodrigues, & Ferreira, 2018). The ability of scientific reasoning promotes students' to acquire domain-specific knowledge, long-term academic achievement, cognitive ability, decision making, and solve daily problems (Andersen & Garcia-Mila, 2017; Bao et al., 2009; Van der Graaf, Van de Sande, Gijsel, & Segers, 2019). Some researchers argue that the societies that can survive in the midst of currently increasing and uncertain challenges are those who are equipped with disciplinary knowledge, understand science as a social enterprise, think critically and creatively, and evaluate and critique information with valid evidence (Hill, 2008; Kind & Osborne, 2017). These days, this leads nations, standard documents, educators, and researchers to focus on how scientific knowledge is constructed, its nature, and how to promote the skill of scientific reasoning (Ding, 2018; Kind & Osborne, 2017; NRC, 2011).

An important aim of education *globally and* for sub-Saharan African countries, in particular, is to produce citizens who can participate in ensuring sustainable development and poverty reduction (Bilican, Cakiroglu, & Oztekin, 2015; Joshi & Verspoor, 2013). As most sub-Saharan African countries intend to attain a middle-income economy, they need a qualified workforce who is trained scientifically with the skills of the 21st century. The workforce who can sustainably retain such complex tasks is the ones who understand key concepts of science, be able to think critically and reason at higher levels, and understand the modern views of the nature of science (Bilican et al., 2015).

However, there is no common consensus about what scientific reasoning is and how it is to be developed. Lawson's argued that scientific reasoning in nature is hypothetico-deductive and consists of skills such as proportional reasoning, control of variables, probability reasoning, and correlation reasoning (Lawson, 2004; Moore & Rubbo, 2011). This conception was developed based on Piaget's theory of cognitive development that includes two formal reasoning levels (concrete operational and formal operational). The concrete operational level involves appropriate use of logic and formal operational reasoning involves thinking abstractly, reasoning logically, and drawing conclusions from available information (Moore & Rubbo, 2011). Based on this view, scientific reasoning is seen as a domain-general skill. On the other hand, there is also another contention that suggests that scientific reasoning is a process that involves different forms of scientific knowledge progressively (Kind & Osborne, 2017). This view suggests that scientific reasoning is not more than developing scientific knowledge at a higher level. The proponents of the domain-specific view of scientific reasoning contend that human cognition is based on the knowledge which is tied to specific activities in specific contexts and which cannot easily be transferred to different activities or contexts (Fischer et al., 2014; Gelman, 2003). .

In the subsequent sections of this chapter, the way scientific reasoning is conceptualized will be discussed based on relevant literature on learning progression, styles of scientific reasoning, and scientific knowledge which, in turn, enables us to develop an assessment tool to measure scientific reasoning.

2.2 Learning progression, styles of scientific reasoning and scientific knowledge

In this section, the construct scientific reasoning will be discussed in terms of how students' learning progresses from low to high level, different forms of scientific knowledge, and styles of scientific reasoning.

2.2.1 Learning progression and construct map

Learners cannot develop a scientific understanding in single lesson and reasoning progresses also can't happen in a moment. A process in the mind of students has to take place before students understand a concept and develop reasoning. Their minds need the necessary materials to establish a better understanding. Learning is not a moment of practice. In all the classrooms where proper instruction is provided, progression is expected to continuously take place. There is a probable path for students to attain every concept to its best of it (Corcoran, Mosher, & Rogat, 2009). The goal of teaching-learning is to help students to attain this understanding in their minds and help them think for themselves (NRC, 2007). Knowledge is considered to be meaningful when it is justified with reasonable evidence. This brings the concept of learning progression.

Learning progressions in science are empirically-grounded and testable hypotheses about how students' understanding of, and ability to use, core scientific concepts and explanations, and related scientific practices grow and become more sophisticated over time, with appropriate instruction (Corcoran et al., 2009; NRC, 2007). These hypotheses describe the pathways students are likely to follow to master core concepts in science. They are based on research about how students' learning progresses—as opposed to selecting sequences of topics and learning experiences based only on logical analysis of current disciplinary knowledge and on personal experiences in teaching. This learning progression is the thought or idea in the minds of researchers, and that it represents a hypothesis or theory about the students' thoughts that will be examined empirically, eventually in the course of study. Learning progression can be defined as students' cognitive learning growth from a lower cognitive level to a higher level of understanding (like Bloom's taxonomy) (Bloom, Hastings, & Madaus, 1971) or students' ways of reasoning and argumentation based on the structural levels (like Toulmin's 1958 structure of argumentation) (Toulmin, 1958).

Toulmin's structure of argumentation describes students' learning progression based on the levels of argumentation and reasoning such as claim, evidence, reasoning, and rebuttal or evaluation whereas Bloom's taxonomy represents students' cognitive learning progress from knowledge to evaluation. There is a relationship between these two progressions in which both start with students simple knowing in lower levels; and at the highest level, both suggest evaluation (Kind, 2013). At the highest level of argumentation, students use all the components of argumentation to arrive at their conclusion. In addition to this, both Bloom's taxonomy, and argumentation and reasoning consist of different forms of knowledge: from content to evidence evaluation (epistemic) which increases its complexity progressively. Epistemology is the highest form of knowledge in which it consists of identifying, evaluating, judging based on ample and justifiable reasons to arrive at a certain conclusion. To arrive at a conclusion students use different explanations and identifying weak and strong alternatives, criticize weak alternatives using rebuttals and limiting boundaries for their claim, and putting things in perspective. One can, therefore, represent the construct scientific reasoning by considering three forms of scientific knowledge (content, procedural and epistemic) in a combined approach and this can also be demonstrated using a construct map instead of dealing with each level of understanding independently.

According to Wilson (2009), construct map is somewhat less complex concept than a learning progression, and is designed to help conceptualize how assessments can be constructed to relate to theories of cognition. The construct map is a well-thought-out ordering of qualitatively different levels of performance in a domain (Wilson, 2005).

As will be discussed in later sections, the current assessment approach (particularly in sub-Saharan African countries) mainly focuses on a single measurement at some final or supposedly significant time point without considering students' learning as a progressive development. According to Wilson (2009) however, this cannot help us to know clearly what to assess and how to assess it, which level of reasoning is attained or not, whether to focus on generalized learning goals or domain-specific knowledge, and the implications of a variety of teaching and learning theories. This is because that as learning situations vary, and their goals and philosophical underpinnings take different forms, a ''one-size-fits-all'' development assessment approach rarely satisfies educational needs. Therefore, a developmental perspective regarding student learning that assesses the development of student understanding of particular concepts and skills overtime is required Wilson (2009). In this regard, he proposed the BEAR (Berkeley Evaluation and Assessment Research) Assessment System which was aimed to provide tools to model many different kinds of learning theories and learning domains. The model was devised to facilitate assessing the growth of students' understanding of particular concepts and skills over a certain period of (instructional) time. It means that it helps one to move away from ''one-shot'' testing situations toward an approach that focuses on the process of learning and an individual's progress through that process. In turn, this provides clear definitions of what students are expected to learn, and a theoretical framework of how that learning is expected to reveal that is also necessary to establish the construct validity of an assessment system. However, such students' learning needs to be mapped in a well orderly sequenced order as indicated in Table 2.1.

To study learners understanding, a construct map need to be established which represents students' thinking based on some expectations. These expectations can be drawn out from a standard document, cognitive theory, and research results. Alternative conceptions can also play a great role in defining useful levels of the underlying construct. The students' alternative conceptions are not considered to be negative rather they are considered as "threshold concepts," which are posited to be transformative, probably irreversible, integrative, possibly bounded, and troublesome (Meyer & Land, 2003). They help in identifying the bottom levels as a result one can easily judge the top levels in consideration of this and standards. In addition to defining student understanding at each level of the continuum, the notion of common errors helps to clarify the difference between levels. Alternative conceptions, represented as common errors in one level, are resolved in the next level of the construct map (Wilson, 2009).

Table 2.1: Construct map for student understanding of Earth in the solar system (Wilson, 2009)

Level	Description
5 8th grade	Student is able to put the motions of the Earth and Moon into a complete description of motion in the Solar System which explains: The day/night cycle The phases of the moon (including the illumination of the moon by the Sun)The seasons
4 5th grade	Student is able to coordinate apparent and actual motion of objects in the sky. Student knows that • The earth is both orbiting the Sun and orbiting on its axisThe Earth orbits the Sun once per yearThe Earth rotates on its axis once per day, causing the day/night cycle and the appearance that the Sun moves across the skyThe moon orbits the Earth once every 28 days, producing the phases of the moon COMMON ERROR: The phases of the moon are caused by a shadow of the planets, the sun, or the Earth falling on the Moon.
3	Student knows that: • The Earth orbits the Sun • The moon orbits the Erath • The earth rotates on its axis However, student has not put this knowledge together with an understanding of apparent motion to form explanations and may not recognize that the earth is both rotating and orbiting simultaneously. COMMON ERROR: It gets dark at night because the Erath goes around the Sun once a day.
2	Student recognize that: • The Sun appears to move across the sky every day • The observable shape of the Moon changes every 28 days Student may believe that the Sun moves around the Erath COMMON ERROR: All motion in the sky is due to the Earth spinning on its axis COMMON ERROR: The sun travels around the Earth COMMON ERROR: It gets dark at night because the Sun goes the once a day. COMMON ERROR: The Earth is the centre of the universe
1	Student does not recognize the systematic nature of appearance of objects in the sky. Students may not recognize that the Earth is spherical. COMMON ERROR: It gets dark at night because something (e.g., clouds, the atmosphere, “darkness”) covers the Sun. COMMON ERROR: The phases of the moon are caused by clouds covering the Moon. COMMON ERROR: The Sun goes below the Earth at night.
0	No evidence or off-track

Table 2.1 represents a construct map for student understanding of Earth in the solar system. The misconception of level one is resolved in level two and the misconception

of level two is resolved in level three. This shows that misconceptions are not strange things that cannot be determined rather they are part of students learning progression. Once we identify the misconception of one level, we can handle it by preparing requirements to promote learner to the next level. The top-level in the construct map represents the understanding expected of 8th graders in national standards. In this way, a construct map helps teachers in planning and delivering their lessons based on students' learning progressions. It helps a teacher to plan a lesson based on evidence of students' understanding of a construct. The teacher who uses students' evidence is not following a haphazard journey in his/her teaching.

However, the construct map depicted in Table 2.1 represents only one style of reasoning, model-based reasoning. It is not satisfactory to focus on only a single style of reasoning of students' understanding that grows progressively. Students' learning and reasoning can grow progressively in terms of various styles of scientific reasoning such as model-based reasoning, data-based reasoning, and etc. There is a need to consider various styles of reasoning in classroom instructions and assessment development to bring meaningful and authentic developments of students' learning. The next section will discuss various styles of scientific reasoning developed in the history of science education.

2.2.2 Styles of scientific reasoning and construct maps

Since science commenced as a school subject and science educators divided the science curriculum into the scientific method and science knowledge, there has been a continuous development of curriculum models (Kind & Osborne, 2017). The two competing models that dominate standards reforms are science competencies and science practices (Kind, 2013). According to Kind and Osborne (2017), both models have conceptual weaknesses related to combining cultural and cognitive perspectives on the science learning domain. The goal of scientific reasoning is to answer three specific questions about the material world (Kind & Osborne, 2017; NRC, 2012). These are the questions of What exists? (the ontic question); Why it happens? (the causal question); and How do we know? (the epistemic question). To answer these questions, science education needs to consider different styles of scientific reasoning which involve the forms of scientific knowledge. Crombie (1994) and Hacking (2012) identified six styles of scientific reasoning. Mathematical reasoning: This style is the use of mathematics to represent the world and for deductive argument. The representation of physical phenomena by mathematics is key to all the sciences. It is taken as the language for physical science. Example: rate change, lines, and geometric shapes. Experimental reasoning: this reasoning aimed to test the predictions and to establish what exists. It is used to test ideas in science whether they are true or not. Ideas which such tests considered to be true otherwise they will be rejected. This can be done by observation and measurement. Example: experimentation in lab class. Conceptual reasoning: This deals with the theoretical representation of the physical world. It suggests that it is possible to represent objects of the natural world that are not accessible to observe and touch clearly to give explanations based on the construction of the model. It is used to represent things that are too large to imagine (the Solar System), or too small to see such as the cell or the atom. It was developed as a method to explain the unknown properties of a natural phenomenon by simulating the phenomenon with the known properties. Example: Wave model of light (to explain diffraction and refraction), we use ray diagrams to reason about the behavior of lenses, or free body diagrams to reason about the effects of forces. Taxonomic reasoning: This reasoning is for the ordering of variety by comparison and taxonomy. It answers the question 'what exists?' it helps in establishing what entities exist. It also serves to establish fundamental similarities and differences. Example: distinguishing concepts such as heat and temperature, mass and weight, and energy and power, magnetic and nonmagnetic materials. Probabilistic Reasoning: It is to predict patterns with a degree of certainty/uncertainty what might happen (to what extent, how often, to what degree some pattern happens).Example: Gaussian distribution, uncertainty principle, wave and particle nature of matter require such reasoning. Historical: This tries to answer the question 'how or why a particular phenomenon comes to be or behave as it does'. It starts with assumptions about how things stand. Attempts to explain the origins of the material world and its features are a major element of reasoning in the sciences.

Students' learning progresses for each style of scientific reasoning from simple to complex form. There are forms of knowledge which represent the lower, middle, and higher levels of reasoning. Therefore, each style of scientific reasoning needs to be defined in terms of domain-specific scientific knowledge and the forms of scientific knowledge need to be identified in each style of scientific reasoning. The next section will discuss about the three forms of scientific knowledge in relation to various styles of scientific reasoning.

2.2.3 Scientific knowledge

Scientific reasoning is dependent on three forms of knowledge: content knowledge, procedural knowledge, and epistemic knowledge (Kind & Osborne, 2017). Content knowledge is the traditionally dominant type of knowledge science teaching has focused on over the years. It helps to answer the ontic question “what exists’’? It is content knowledge appropriate to domain-specific concepts—that is the ontological entities that science uses to reason with. According to OECD (2017), content knowledge involves knowledge of the facts, concepts, ideas, and theories about the natural world and the explanations. Procedural knowledge is knowledge of the procedures and associated constructs that science uses to establish its claims to know. It helps to answer the causal question “Why it happens?’’ Procedural knowledge is related to scientific evidence to accept or reject a claim (Gott, Duggan, & Roberts, 2008). In model-based reasoning, for example, scientists use representations such as rays and wave-models to explain sound and light, and free-body diagrams to solve problems in mechanics. Similarly, in data-based reasoning, they use standardized procedures to design experiments, make reliable measurements, and present and analyze data. Epistemic knowledge is the knowledge of the epistemic constructs and values and how these are used to justify science’s claims to know. It is the concepts and ideas used to describe scientific knowledge and methods. Epistemic knowledge involves evaluating claims, explanations, evidence, hypothesis, models, and theories in order to establish scientific knowledge (OECD, 2017). It helps to answer the epistemic question ‘’ How do we know’’? Scientists use concepts such as dependent-, independent- and control-variables in experimental reasoning, and they talk about fair testing and control of variables.

In model-based reasoning, they distinguish between scientific laws, theories, models, and hypotheses. Furthermore, a key issue in all scientific reasoning is the criteria used for evaluating knowledge and data, such as the support in empirical evidence. These criteria have to be learned by novices entering the scientific culture.

In order to develop learning progression from simple to complex levels of reasoning, a construct to be assessed needs to be defined in a way that incorporates various styles of scientific reasoning as well as the three types of scientific knowledge: content, procedural and epistemic. A construct, here, is a variable or an operationalized concept one wants to measure in a particular study. In this study, the construct being addressed to study is scientific reasoning. The consideration of styles of scientific reasoning, scientific knowledge, and learning progression provides a sound and meaningful representation of the construct in assessing students' abilities. Thus the construct, scientific reasoning, is conceptualized in terms of learning progression, styles of scientific reasoning, and different forms of scientific knowledge in the section below.

2.2.4 Conceptualizing scientific reasoning

An extensive review of literature allows identifying what entails to the construct to be studied. Before the creation of items, the construct under study needs to be defined explicitly. For this study, the construct to be studied i.e. scientific reasoning was defined in terms of learning progression and Styles Scientific of Reasoning (SSR). Scientific reasoning is conceptualized and four progressive levels were identified (see section 2.4).

All aspects of human cognition are expressed by processes such as perception, attention, memory, categorization, logical reasoning, and associative learning considered as domain-general (Gelman & Noles, 2011). Consequently, previous studies of scientific reasoning gave more weight to the view that scientific reasoning is domain-general (Kuhn, 2002; Zeineddin & Abd-El-Khalick, 2010). Based on this assumption scientific reasoning was considered as independent of content knowledge and it was assumed to be merely a skill to be developed. This view led researchers to develop assessments that can measure students' ability of scientific reasoning and skills such as control of variables, generating hypotheses, generating evidences, evaluating evidence, and drawing conclusions independently (Kuhn, 2002; Zeineddin & Abd-El-Khalick, 2010) without emphasizing the content knowledge. Recently this view has substantially been challenged by the view that scientific reasoning is dependent on domain-specific knowledge (Kind & Osborne, 2017; Osborne, 2013; Zeineddin & Abd-El-Khalick, 2010). Studies revealed that children construct domain-specific theories which, in turn, lead to domain-specific reasoning from infancy through adulthood (Gelman & Noles, 2011). Moreover, pre-school children and elementary school students have been found to be engaged in reasoning and higher-level activities using domain-specific knowledge (Gelman, 2003; Gelman & Noles, 2011). Studies also reveal that human cognition such as students' reasoning ability is highly related to domain-specific knowledge and has a gradual developmental pattern starting from an early age (Gelman & Noles, 2011). It is essential to address the primary school students' domain-specific scientific reasoning ability because such an age is characterized by grasping the basics of physics knowledge.

These noted that it is important to assess not only competencies of reasoning skills but also the three types of knowledge: content, procedural and epistemic in relation to styles of scientific reasoning and how such skills can be manifested in real-life situations.

To this end, recent studies give attention to relating styles of scientific reasoning to the competencies of reasoning instead of dealing with competencies of reasoning alone (Kind & Osborne, 2017; Osborne et al., 2018). Kind and Osborne (2017) argued that giving much focus on reasoning skills alone would not help to bring a sustainable improvement in the whole picture of students' learning progress. Styles of reasoning are advantageous over the other assessments of scientific reasoning frameworks in that it recognizes the need for all the three elements of domain-specific knowledge and different forms of reasoning that science has developed over the years. Even though thinking skills are transferrable, the ability to think and reason critically requires in-depth knowledge and understanding of a particular domain and is dependent on domain-specific content knowledge (Davies, 2013; Tiruneh, De Cock, Weldeslassie, Elen, & Janssen, 2017). Scientific reasoning, therefore, in this study is considered as a subject knowledge-dependent, domain-specific, and is expressed through learning progression and styles of scientific reasoning. Scientific reasoning here defined by combining the frameworks of Ford and Wargo (2012) and McNeill and Krajcik (2011) in a way to contain the three types of scientific knowledge (level 1 and 2 – content knowledge, level 3 procedural knowledge, and level 4 epistemic knowledge) for the three styles of scientific reasoning.

The conceptualization of the construct, scientific reasoning, is based on the review of relevant literature. In addition to this, the development of assessment tools that consist of styles of scientific reasoning, scientific knowledge, and learning progression need also be guided by research frameworks that enable to represent a construct to be measured in different levels of reasoning. The following section will present the frameworks in learning progression and scientific reasoning.

2.3 Research frameworks for learning progressions and scientific reasoning

Various theoretical frameworks can help to develop the learning progression of students' scientific reasoning ability. However, most frameworks are not developed by incorporating all the forms of knowledge and one or two forms of knowledge in the progressive reasoning levels. Some frameworks lack progressive reasoning levels in terms of content knowledge explicitly (for example (McNeill & Krajcik, 2011); others lacks to incorporate procedural knowledge in their levels of learning progression (for example, Ford

& Wargo, 2012); some other also lack to incorporate content knowledge and progressive reasoning levels in terms of content knowledge (for example Toulmin's framework).

This study intends to develop a conceptual framework that incorporates all the forms of knowledge along with different styles of scientific reasoning. To this end, different frameworks are discussed in a way to identify the strengths and weaknesses to develop the conceptual framework of this study that can guide the development of assessment tools.

2.3.1 Toulmin's (1958) Framework

It is required to have frameworks to assess students' way of reasoning and develop assessment tools. A framework is used to guide in depicting what type of learning to be considered, what types of knowledge need to be included, and what different levels of reasoning represent students' ability.

Toulmin's (1958) framework is considered as a foundational framework for the analysis of arguments by introducing various components. Various frameworks have been developed to assess students' reasoning and argumentation by taking Toulmin's (1958) framework as a foundational framework (Sampson & Clark, 2008). It consists of mainly claims, data, warrants, backings, qualifiers, and rebuttals. Claims are statements that are being argued. Data are evidence that supports claims. Warrants are comments that are used to justify why data are relevant to the claims. Backings are statements that serve to support the warrants/strengthen the acceptability of the warrant. Qualifiers explain conditions under which the claim holds. Rebuttals are statements that contradict the data, warrants, or backings. This framework was developed in a way that it can measure arguments based on the structure students provide. According to this framework, the argument which consists of all the components of the structure is considered to be strong and the argument with the lest components of the structure is considered as a weak argument without considering the validity of the content of the argument.

However, Toulmin's framework does not consider the significance of scientific knowledge demonstrated in students' arguments rather structural components (Sampson & Clark, 2008). It does not directly assess the content knowledge, it only focuses on examining whether the argument components are included or not. According to Toulmin's framework, although the student argument contains most components of the argument, the content can be inaccurate and inappropriate from a scientific perspective. For example, students can generate arguments with wrong content knowledge.

2.3.2 McNeill & Krajcik's (2011) framework

McNeill and Krajcik's (2011) framework provides teacher's the necessary guidance to support students in constructing scientific explanations for grades 5–8. This framework is developed by adapting Toulmin's (1958) model of argumentation (McNeill & Krajcik, 2011). The framework consists of four structures: claim, evidence, reason, and rebuttals (see Table 2.2). According to the framework, the claim is a statement that expresses the answer or conclusion to a question or problem; the evidence is scientific data that supports the claim; reasoning involves providing a justification that links the evidence to the claim; and the rebuttal recognizes and describes alternative explanations and provides counter-evidence and reasoning for why the alternative is not the appropriate explanation for the question or problem. This framework has different variations based on students' ability, age difference, and grade level. The first variation involves providing a claim, evidence, and reason. The second variation involves three components but it requires the appropriateness and sufficiency of the evidence. The third variation involves the three components but each piece of evidence required different justifications. The last variation is a complex one for it involves rebuttals which requires students to provide exceptions in which the claim does not work.

The framework considers the explanation as building knowledge of how and why a phenomenon occurs, and argument as socially constructing knowledge when there is a need to justify a claim and persuade others of its validity (Berland & McNeill, 2010). Their framework, however, does not provide different progressive levels of reasoning in terms of content knowledge; it merely focuses on the process of argument. Even if the framework is developed in a way that can incorporate both content and structure of reasoning and argumentation, the levels are categorized based on the mere structure of the argument, not according to progressive content knowledge. This suggests that the focus of the framework is only procedural and epistemic knowledge.

Table 2.2 McNeill &Krajcik (2011) Framework for scientific explanation

Framework Sequence	**Description of Framework for Students**
Claim	*Claim* • A statement that answers the question
Evidence	*Evidence* • Scientific data that support the claim • Includes multiple pieces of data
Reasoning	*Reasoning* 1. A justification for why the evidence supports the claim using scientific c principles
Rebuttal	*Rebuttal* 2. Describes alternative explanations, and provides counterevidence and counter reasoning for why the alternative explanation is not appropriate

2.3.3 Ford and Wargo Scaffolding Framework

Ford and Wargo (2012) identify four science epistemic practices with increasing cognitive demand (see Table 2.3). This framework describes how students' content knowledge understanding progresses from low level (recounting) to high level (evaluation). The first level of the hierarchy is mere recounting generally (i.e., without reference to data) what an explanatory scheme is. At this level students are expected to construct models consistent with prior evidence and theories to illustrate, explain, or predict phenomena. The second level is an application of an explanatory scheme to explain a particular piece of data or a natural phenomenon. In this model, students use models to illustrate, explain, and predict phenomena. In mathematical-deductive reasoning, they use formulas to solve physics problems. That is, students should recount scientific knowledge, in addition to this, they use the knowledge to explain phenomena in nature and solve scientific problems. The third level is a juxtaposition of more than one explanatory scheme. At this level, students compare and evaluate the ability of different models to accurately represent and account for patterns in phenomena, and to predict new phenomena. This involves explaining the design in a scientific experiment, identifying the line of argument in a scientific explanation, or separating the model from the science phenomena it aims to explain. The fourth level is the evaluation of an explanatory scheme in light of its ability to account for data or a natural phenomenon, relative to another scheme. Here, students are expected to revise models to increase their explanatory and predictive power, taking into account additional evidence or aspects of a phenomenon. This practice primarily regards critique and evaluation and it aims at validating knowledge, methods, and arguments.

The Ford and Wargo (2012)'s framework can help one to assist students in classroom discourse identifying various levels of reasoning. However, the framework does not provide the way one can assess the quality of each level of reasoning explicitly. It does not provide measuring each level with the structures of the argument. This indicates that the framework focuses on the content and epistemic knowledge in developing the learning progression. It lacks procedural knowledge which is known by students providing evidence for the claim they generated.

Table 2.3 Ford and Wargo (2012) Framework

Levels	Operations
Evaluate	Evaluate the explanatory Scheme
Juxtapose	Consider two or more explanatory schemes simultaneously
Explain	Explain a specific phenomenon using a particular explanatory scheme
Recount	A general restatement or clarification of an explanatory scheme
Nonact	A statement that does not connect to any explanatory scheme

From Section 2.2 and 2.3, one can see that to assess and develop assessment tools a construct needs to be conceptualized (see Section 2.2.4) by considering three forms of scientific knowledge (see Section 2.2.2) and Styles of Scientific Reasoning (see Section 2.2.3) that is guided by a suitable theoretical framework which depicts student learning progressions (see Section 2.2.1). The following section will discuss how the conceptual framework is developed that guides the assessment of scientific reasoning and the tool that represents different progressive reasoning levels.

2.4 Conceptual framework

It was suggested that classroom instruction and assessment development procedures need to address three forms of scientific knowledge, various styles of reasoning, learning progression, and evidence-based approaches (Corcoran et al., 2009; Kennedy & Wilson, 2007; Kind & Osborne, 2017; Wilson, 2009). To address such issues this study approached the construct scientific reasoning from two-dimensional structures: one Styles of Scientific Reasoning (SSR) and the other proficiency levels. The Styles of Scientific Reasoning framework (see Kind and Osborne (2017) includes six styles of scientific reasoning identified through analysis of the history of science: mathematical deduction, experimental evaluation, hypothetical modeling, categorization and classification, probabilistic reasoning, and history-based evolutionary reasoning. These styles of scientific reasoning recognize the need for three elements of domain-specific knowledge: content, procedural and epistemic (Kind & Osborne, 2017). Among the six SSR, the first three styles are commonly represented and dominated in primary school science (Hacking, 2012). Reasoning in primary school physics may require mathematical-deductive reasoning, model-based reasoning, and data-based (Alemu et al., 2017). In Model-based reasoning, students may use scientific concepts, theories, laws, principles, and relationships to provide a claim, generate evidence, and draw conclusions. In data-based reasoning, students may use data in tables and graphs to provide a claim, generate evidence, and draw conclusions. In mathematical reasoning, students may use mathematical formulae, variable relationships to provide a claim, generate evidence, and draw conclusions. Hence, this study framed the scientific reasoning construct on these three SSR and developed the SRPT that assesses grade 8 students' scientific reasoning ability.

Literature confirms that students' scientific knowledge and reasoning abilities progress from lower to higher reasoning levels based on the currently discovered knowledge and understandings (Berland & McNeill, 2010; Ford & Wargo, 2012; Kind, 2013; McNeill & Krajcik, 2011). Besides, these studies represent a learning progression in a construct map by identifying what students can and can't do in each level of progress along with possible ways of progression (Alonzo & Steedle, 2009; Berland & McNeill, 2010; Corcoran et al., 2009; Kennedy & Wilson, 2007; Wilson, 2009). This indicates that proficiency levels require identifying the progress of students' understanding and reasoning from tacit towards a higher form of understanding and reasoning. Proficiency levels for this study are identified based on the Ford and Wargo (2012) and McNeill and Krajcik (2011) theoretical frameworks.

The Ford and Wargo (2012) framework is developed to address conceptual understanding and epistemic knowledge from recount to evaluation level whereas McNeill and Krajcik's (2011) framework was developed in a way to incorporate the structure of reasoning. The Ford and Wargo (2012) framework gives more focus on the content and epistemic knowledge. They developed a scaffolding framework that shows students learning progression in a way to answer what and how students' knowledge and understanding ability progresses from low to a high level. The framework gives less emphasis on the structure of the reasoning, particularly evidence generation, which describes students' procedural knowledge. On the other hand, McNeill and Krajcik's (2011) framework represents the structure of reasoning from claim to rebuttal which tends to describe students' procedural and epistemic knowledge. In addition to this, the Ford and Wargo (2012) framework informs students' progress of conceptual understanding but it gives less emphasis towards students' pattern of reasoning. The McNeill and Krajcik (2011) framework considers the structure of reasoning which also gives less emphasis to conceptual understanding.

Therefore, the combination of both frameworks enabled the development of a conceptual framework which guides the development of scientific reasoning learning progress levels and assessment tool which considers:

- Three types of knowledge (content, procedural and epistemic);
- Progressive learning (define proficiency levels);
- Construct-based (conceptualization of the construct); and
- Domain specific (knowledge-dependent scientific reasoning).

Table 2.4: The proficiency levels of McNeill & Krajcik (2011), Ford and Wargo (2012), and conceptual frameworks

<table>
<tr><th>Levels</th><th colspan="2">Proficiency levels of McNeill & Krajcik (2011) framework</th><th colspan="2">Proficiency levels of Ford and Wargo (2012) framework</th><th colspan="2">Proficiency levels of conceptual framework (for this study)</th></tr>
<tr><td>1</td><td>Claim</td><td rowspan="4">Involves only content and epistemic knowledge; lacks procedural knowledge explicitly</td><td>Recount</td><td rowspan="4">Involves procedural and epistemic knowledge; lacks content knowledge and LP in terms of content knowledge</td><td>Factual knowledge</td><td rowspan="4">Involves content, procedural and epistemic knowledge; LP is in terms of three forms of knowledge</td></tr>
<tr><td>2</td><td>Evidence</td><td>Explain</td><td>Explanation of factual knowledge</td></tr>
<tr><td>3</td><td>Reason</td><td>Juxtapose</td><td>Evidence-based reasoning with juxtaposed knowledge</td></tr>
<tr><td>4</td><td>Rebuttal</td><td>Evaluation</td><td>Drawing conclusions based on critical evaluation.</td></tr>
</table>

Based on the amalgamation of these two frameworks, the proficiency levels that depict students' progress of scientific reasoning were depicted in Table 2.4. The first level of McNeill and Krajcik's (2011), a claim can be categorized under the first and the second level of Ford and Wargo's (2012) because both recount and explanation are about students' claims in different levels of understanding. Therefore, the first level of the conceptual framework is the generation of factual level which demands students recounting knowledge based on simple facts and the second level is the explanation of claim by relating and integrating factual knowledge. The first and the second level of this conceptual framework represent content knowledge because according to OECD (2017), this form of knowledge involves the facts, concepts, ideas, and theories about the natural world with its explanation. The third level is evidence-based reasoning based on McNeill & Krajcik (2011) that can obtain students' evidence for the explanations they provided for the second level and the evidence is based on the juxtaposed knowledge according to Ford and Wargo (2012). The third level of this conceptual framework, evidence generation, represents procedural knowledge because the evidence is considered to be procedural knowledge (Gott et al., 2008). The fourth level is drawing-conclusion by considering the high-level of understanding according to Ford and Wargo (2012) and by evaluating available evidence according to McNeill and Krajcik's (2011) framework. The fourth level of this conceptual framework is considered as epistemic knowledge because it is considered as establishing scientific knowledge by using procedural knowledge (OECD, 2017).

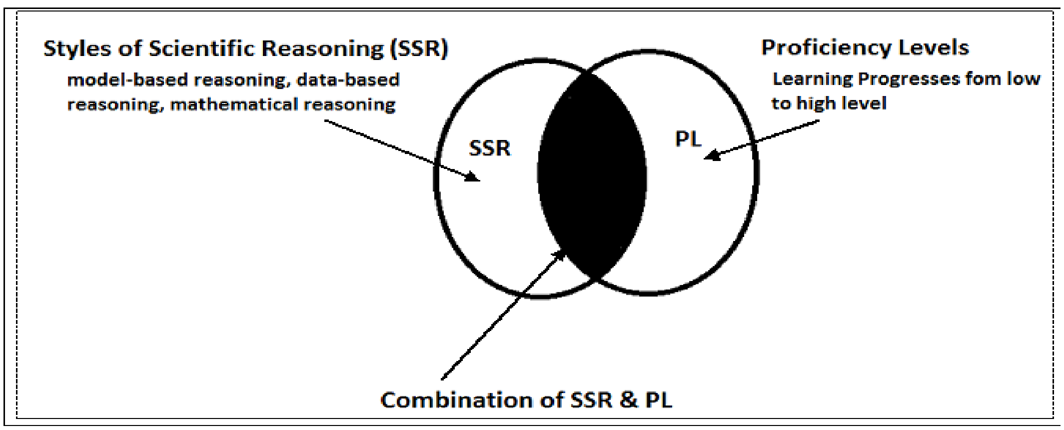


Figure 2.1: Conceptual framework for the study which relates scientific styles of reasoning and proficiency levels.

The conceptual framework in Figure 2.1 demonstarates the combination of styles of scientific reasoning and proficiency levels that are depicted in Table 2.5 explicitly. This combination provides the inclusion of the three forms of scientific knowledge, three styles of scientific reasoning, and students' reasoning progresses from factual knowledge to drawing conclusion. The conceptual framework demonstrated in Table 2.5 explicitly guided the development of the item map demonstrated in Table 4.10 which, in turn, guided the development of the scientific reasoning progress tool along with the person map demonstrated in Table 4.12.

Table 2.5 The demonstration of conceptual framework by combining proficiency levels with styles of scientific reasoning

		Styles of Scientific Reasoning		
		Mathematical deductive reasoning	**Model-based reasoning**	**Data-based reasoning**
Levels of Proficiency	**Drawing conclusion**	Epistemic knowledge	Epistemic knowledge	Epistemic knowledge
	Evidence-based reasoning	Procedural knowledge	Procedural knowledge	Procedural knowledge
	Explanations of claim	Content knowledge	Content knowledge	Content knowledge
	Factual knowledge	Content knowledge	Content knowledge	Content knowledge

Having establishing conceptualization of the construct being studied in this study and the conceptual framework that guided the development of assessment with review of relevant literature, in the following sections the concept assessment will be discussed in relation to scientific reasoning.

2.5 The alignment between assessment, curriculum and teaching

Assessment of student learning provides evidence so that educational decisions can be made (Corcoran et al., 2009; Goldman & Pellegrino, 2015). Any decision that is made by the evidence of assessment leads to constructive progress. In contrary to this, any decision that is made without consideration of valid assessment leads to a confusing result that cannot help to point out what to do and how to go forward. It is difficult to decide without valid information. The assessment helps to identify strengths and weaknesses; it also suggests what might be done to reinforce the strengths and to overcome the weaknesses. To improve students' learning, clear identification between theory and practice has to be identified. Implementing theory without considering students' practical experience yields a vague result that cannot be measured clearly. This can also bring difficulty in identifying where the problem is; whether in the learner, in the theory, or the practice itself. To bring fruitful change, we need to identify what works and what does not work in the context of the students. We need to answer the question ''what does theory on students cognitive structure state on a particular construct?'' and ''what can and cannot our students do practically?''. Before implementing a new model or learning theory it is of paramount importance to give priority to knowing our students concerning the method or theory to be implemented. Knowing our students help us to identify what to teach, how to teach, and what instructional materials are appropriate for students learning (Goldman & Pellegrino, 2015). It is valuable for educational management to judge or make a decision about teaching-learning. If this is so, the question that follows is, ''how can this be done?'' This is to stress how much educational assessment is powerful in bringing change and improvement in teaching-learning. Assessment is a means to identify what is working and what is not working. It tells us what is missed and what is needed in teaching-learning. Assessment should mirror good instruction; happen continuously as part of instruction; and provide information (to teachers, students, and parents) about the levels of understanding that students are reaching (NRC, 2000). Assessment of student learning provides evidence so that educational decisions can be made by different bodies. One can use the evidence to evaluate a teaching program or to make statements about student competence or to make decisions about the next aspect of teaching for particular students.

For effective teaching-learning to occur there has to be a better alignment between assessment, curriculum, and instruction. Curriculum, assessment, and instruction need to support each other to improve students' learning. According to NRC (2001), to support students' classroom learning the educational assessments need to be developed in a way that can align with curriculum and instruction. In addition to this, Martone and Sireci (2009) argue that, even though assessments, standards, and instruction have each been determined and enacted at multiple levels of the educational structure, they are all integral to students' achievement. As they argue further stating content standards (embodied in state curriculum frameworks) represent state-level policy documents, but the policymakers do not create the assessments, and the curriculum standards and assessments are implemented at the local level. Clearly stated objective shows direction what students are expected to learn. This objective can be achieved via instruction. Assessments are required to show explicitly how well students have learned what we expect them to learn. This link help to reveal the overall act of teaching-learning, in turn, promote better students learning progress. Contrary to this, current assessments provide very limited information for teachers and educational administrators to identify why students do not perform well or to modify the conditions of instruction in ways likely to improve student achievement (NRC, 2001). In addition to this, they provide "snapshots" of achievement at particular points in time, but they do not capture the progression of students' conceptual understanding over time. That is, there is no clear alignment between current assessments and instruction. Assessment is required to give information for students, teachers, and other stakeholders so as to improve students' learning and inform proper instructional decisions.

There has to be a clear objective that students are expected to learn. Assessments are required to show explicitly how well students have learned what we expect them to learn. Through instruction, this can be achieved in a classroom. This shows that alignment between assessment, curriculum, and classroom instruction has a crucial aspect in improving teaching-learning. For effective students' learning progress one needs to bring the three aspects in a proper context.

2.6 Assessment in Education

Assessments in education require determining how well students are learning (NRC, 2001). It is a significant part of education in that it provides feedback to students, educators, parents, policymakers, and the public about the effectiveness of educational learning. According to NRC (2001), educational assessments are playing a greater role in decision making than enhancing instruction and classroom learning. It is required to emphasize classroom assessment than assessment for decision making to support learners in constructing their knowledge. Classroom assessments are activities undertaken by teachers and by their students in assessing themselves to provide information as feedback to modify teaching and learning activities (Black & Wiliam, 1998). Assessment for learning is the process of seeking and interpreting evidence for use by teachers and learners to decide where learners are, where they need to go and how best to get there (Arega (2014). This can be achieved by using strategies such as questioning, observing, discussing, checking on students' understanding, and analyzing their responses on a regular basis. In addition to assessment for learning, assessment of learning is also used to confirm what students know and can do, to demonstrate whether they have achieved the expected learning outcomes. This can be used to ensure what students know and can do at the end of the unit, course, term, semester, or year. To support students learning it is required to have assessments of students' ability and assessment instruments that identify their levels of progress indicating how students are likely to think and what they are likely to know, understand, and be able to do along with their likely misunderstandings as well (Corcoran et al., 2009).

There are different assessment models in education: Item-driven model and construct-driven model. Item-driven models can also be categorized into criteria-based and content-based. The criterion-based model is the development of assessment based on stated criteria. It addresses the question of how well a test estimates the criterion, which could be defined in terms of performances of the actual task (Kane, 2006). In such a model, the focus is to pass or fail students based on the stated criteria. Teachers might focus on attaining criteria instead of assessing the authenticity of learning. On the other hand, the content-based model is based on the stated students' understanding level and the content involved in the program. The main focus is attempting to cover the contents students are expected to learn; still, much attention is not given for authentic learning. Teachers strive to make sure that their students achieve the expected outcome. Content-based assessments can provide an adequate analysis of validity for observable aspects or operationally defined attributes but are not adequate for theoretical (hidden) constructs (Sireci, 2007). Both content-based and criteria-based are categorized under the item-driven assessment, which tends to omit the construct definition and focus more directly on available items, for example, in an already established item bank. If the item statistics are satisfactory, the models assume the instrument is considered finished or validated (Brown & Wilson, 2011). They give a conclusion from a score point of view; they don't tell us the level of students understanding concerning a specific construct. One of the most serious problems in assessment is not only using statistical methods for analyzing items but also simply that test development has been far too item-driven (Corcoran et al., 2009). The major aim of the item-driven assessment is developing and validating items and analysis of psychometric properties.

In education, it is particularly important that tests support good teaching and learning, and that tests assess what educators think is important. Assessors, therefore, have also been striving to find better procedures for "construct-driven" and "evidence-based" assessment and to developing assessments that support learning progression (Kind, 2013; Wilson, 2009). Unlike item-driven assessment, construct driven model provides "maps" that enable teachers to tell where students are in a learning journey, not just to see whether they have arrived or fallen short, but rather to help them see where to go next or what has been missed (Corcoran et al., 2009). It tells the level of student understanding in relation to the expected learning goals. It is an argument-based approach to validation; it presupposes that it is the interpretations and uses of test scores that are to be evaluated (Kane, 2006). It does not only focus on the test but the test and the proposed interpretations and uses of the test scores. The major aim of construct-based is not developing and validating the items rather analyzing current scientific knowledge, identifying what students can do and can't do (identifying different levels of students understanding or reasoning), identifying the possible misconceptions, and establishing a framework to enhance authentic and meaningful classroom learning.

In this regard, there is a need for properly devised assessments of students reasoning ability and tools in science education which measure such abilities which guide teachers to develop classroom instruction, to identify students' ability level, to identify alternative conceptions, to help students progress from one level to another and to make students lifelong learner by themselves.

2.7 Trends of Global assessments

Globally there are several high stake tests and large scale international assessments that are meant to provide comparative evidence on the extent to which students have mastered and apply their knowledge of school curriculum content in different disciplines, to examine shared students' achievement at an international level, to assess the ability of students concerning common core elements of curricular objectives and to assess skills which characterize among the nations (Au, 2007; Cresswell et al., 2015; Kind, 2013). These assessments are aimed to assess whether the basic needs of children meet, to improve quality of education, to help in making decisions about students, educators, administrators, schools, and communities (Au, 2007; Cresswell et al., 2015). In this regard, the well-known and dominant assessments such as the Lawson classroom test of scientific reasoning, TIMSS, PISA, PIRL, and NAEP are discussed in the following section. These assessments are dominant, well-established for a long time, and commonly used items in science education (Bao et al., 2009; DeBoer, 2011; Linn & Baker, 1995).

2.7.1 Lawson classroom test of scientific reasoning

Lawson's Classroom Test of Scientific Reasoning (LCTSR) was developed based on Piaget's (1954) formal operational schemata which consider scientific reasoning as hypothetico-deductive and consist of skills such as proportional reasoning, control of variables, probability reasoning, and correlation reasoning (Lawson, 2004; Moore & Rubbo, 2011). Piaget's theory of cognitive development includes classification into two formal reasoning levels (concrete operational and formal operational). The concrete operational level involves appropriate use of logic and formal operational reasoning involves thinking abstractly, reasoning logically, and drawing conclusions from available information (Moore & Rubbo, 2011). However, the Lawson classroom test was heavily criticized in that it focuses on a domain-general aspect of cognitive thinking (Bao et al., 2009; Kind, 2013; Osborne, 2013). The proponents of domain-specific view of scientific reasoning contend that human cognition is based on the knowledge which is tied to specific activities in specific contexts and which cannot easily be transferred to different activities or contexts (Fischer, Chinn, Engelmann, & Osborne, 2018; Gelman & Noles, 2011).

2.7.2 Assessment of TIMSS

The Trends in International Mathematics and Science Study (TIMSS) is also one of the international assessment projects that have been engaged in developing items in mathematics and science achievement of students for the fourth and eighth-grade students.

TIMSS 2015 organized the science assessment framework into two dimensions: The content dimension, specifying the subject matter to be assessed; and the cognitive dimension, specifying the thinking processes to be assessed. The TIMSS content domain includes biology, chemistry, physics, and environmental science for grade eight students' and the cognitive aspects include knowing, applying, and reasoning (Mullis, Martin, & Sainsbury, 2016). The three cognitive aspects were treated as an independent entity and reasoning was considered as a domain-general aspect. They developed items for each cognitive aspect independently without relating these aspects in a manner that the skills depend on each other and in ways to demonstrate students' learning develops progressively.

2.7.3 Assessment of PISA

The Program for International Student Assessment (PISA) aimed to assess three basic academic competencies: reading comprehension, mathematical competence, and scientific competence. According to OECD (2017), the science assessment entails scientific literacy that comprises of competencies in science and technology such as:

- Explain phenomena scientifically – recognize, offer, and evaluate explanations for a range of natural and technological phenomena.
- Interpret data and evidence scientifically – analyze and evaluate data, claims, and arguments in a variety of representations and draw appropriate scientific conclusions.
- Evaluate and design scientific inquiry – describe and appraise scientific investigations and propose ways of addressing questions scientifically.

One of the important aspects PISA underscored was inquiry-based science education (IBSE) (Sjoberg, 2019). This is because IBSE was expected to promote students' critical thinking, gain students' interest and motivation to learn scientific knowledge and provide a positive attitude towards science learning. Furthermore, PISA conceptualization of the construct i.e. the scientific competency is based on the three types of knowledge: content, procedural, and epistemic. However, it is yet criticized for lack of explanation of the link between knowledge dimensions and scientific reasoning (Kind, 2013). Though scientific reasoning and critical thinking were some of the concepts addressed by PISA assessment in the form of IBSE and scientific competency, they are embedded in the construct scientific competency. The construct scientific reasoning is not explicitly conceptualized using domain-specific scientific knowledge. That means the framework unable to create a link between the three forms of scientific knowledge and lacks to develop a progression for each form of scientific knowledge to conceptualize scientific reasoning.

2.7.4 Assessments of NAEP

The National Assessment of Educational Progress (NAEP) an American assessment framework that is engaged in assessing grade 4, 8, and 12 students to determine what students know and do in science mathematics, and other subjects. Though this framework is an American assessment framework, the focus given for the domain-specific scientific knowledge (especially for the procedural and epistemic) has a significant contribution to the development of assessment tools. The NAEP framework defines science content in terms of three broad fields of science: physical science, life science, and earth science (see Table 2.6). NAEP categorizes the competencies of science into conceptual understanding, scientific investigation, and practical reasoning. NAEP measures students' performance based on three achievement levels (low, moderate, and high) for each age and grade in each subject area to be tested (NAGB, 2015). The development of the levels considered to be consistent with relevant widely accepted professional assessment standards and based on the appropriate level of subject matter knowledge for grade levels to be assessed, or the age of the students. The levels are:

- Basic: This level represents a partial mastery of prerequisite knowledge and skills that are fundamental for the next level.
- Proficient: This level indicates a solid academic performance for each grade assessed. Students who reach this level are characterized by demonstrating competency over a difficult subject matter such as mastering subject-matter knowledge, applying such knowledge to real-life situations, and developing analytical skills..
- Advanced: This level implies superior performance beyond the level of Proficient

The first level is characterized by describing and explaining nature as it is and the second level is characterized by students' awareness of multiple alternatives for a science concept and the final level regarded as students use of evidence either to accept or reject among the multiple explanations.

Table 2.6 The NAEP 2015 science framework

		Science content		
		Physical science Content statements	Life science Content statements	Earth and space sciences Content statements
Science practices	Identifying science principle	Performance expectation	Performance expectation	Performance expectation
	Using science principle	Performance expectation	Performance expectation	Performance expectation
	Using scientific inquiry	Performance expectation	Performance expectation	Performance expectation
	Using technological design	Performance expectation	Performance expectation	Performance expectation

The NAEP framework has strengths in incorporating current perspectives in learning science and science studies; however, it is criticized in that it lacks explicit categorization of the knowledge and reasoning dimensions explicitly (Kind, 2013). The NAEP framework tried to address scientific reasoning using domain-specific procedural and epistemic knowledge, the construct scientific reasoning is embedded in the construct science practices that comprise of asking questions and defining problems, developing and using models, planning and carrying investigations, analyzing and interpreting data, using mathematics and computational thinking, constructing explanations and designing solutions, engaging in argument from evidence, and obtaining, evaluating and communicating information (Sjøberg, 2018). This indicates that it lacks the conceptualization of the skills (such as scientific reasoning) being assessed by giving explicit emphasis to scientific knowledge.

The tests discussed above have contributed to the progress of students' learning for decades. They have had significance in holding teachers accountable for students' learning to achieve what students are expected to learn, to achieve the quality of education, to set high expectations that suits globally accepted skills, to provide students' information for the concerning body, and to compare students' results among nations (Cresswell et al., 2015). The test results have made an impact on policy design, curriculum revisions, and promotions of students, educators, and schools. These assessments have had also great importance in developing and improving educational policy, curriculum development, educational practices, improved learning experiences, and positive learning outcomes among participating countries.

However, such tests are criticized for an inability to contextualize current educational theory towards the needs of learners and societies, ways of learning, and available resources; they faced challenges in conceptualizing various learning domains by creating a meaningful link between domains of learning; they faced difficulty in devising assessment procedure that suits all the countries (Au, 2007; Creswell et al., 2015). The trends of the assessments might limit teachers to focus instructions on the content areas that the tests frequently appear and to prepare students for the exam instead of helping students' in developing higher-level reasoning and knowledge construction.

They are also criticized for narrowing curriculum objectives and lack of incorporating minority groups adequately (Au, 2007). The tests have also faced difficulty in conceptualizing the construct to be assessed. A single test can't be expected to conceptualize various domains of learning into one construct. This is evidently seen in large scale assessments for the position they hold in attempting to assess the construct, scientific reasoning (Kind, 2013). Some assessments try to develop tests for scientific reasoning as a separate element from content knowledge and some others don't put clear distinction and similarity of scientific reasoning from different types of knowledge.

This indicates that there is a need to have domain-specific assessments of scientific reasoning that consider contextual factors and that conceptualize scientific reasoning in terms of three types of scientific knowledge.

2.8 Assessment in low-income nations

In assessing the abilities of students and developing items there are issues that need to be considered such as suitability of the items for the students in terms of ability level, the difficulty of items, socio-cultural background, construct to be measured, and forms of knowledge involved. Students in low-income nations lack facilities like television, internet, equipped laboratory classes, and even relevant reference materials. We cannot compare students without considering who has the first-hand experience of some phenomena that enhances understanding and can give some students an advantage over others. Therefore, there has to be such consideration when items are developed.

Research findings show that primary and secondary school students in most developing countries learn less than (in terms of content, the duration they elapsed with the lessons, materials provided, and skilled and knowledgeable teachers) their counterparts in developed countries (OECD, 2011). Studies also show that there is a significant achievement gap between high-income nations and low-income nations (Sabanathan et al., 2015). Some estimates put this gap at about two standard deviations (Chmielewski & Reardon, 2016; Glewwe & Kremer, 2006). The reason identified for this gap included differences in family characteristics and low school quality in developing countries. It can also be poverty, malnutrition, recurrent, and/or chronic infectious diseases plus inadequate cognitive stimulation and low school quality in developing countries (Sabanathan et al., 2015). Children's mental and physical growth is dependent on many factors such as biological factors (genes, brain growth, neuromuscular maturation), and environmental influences (parent-child relationships, community characteristics, cultural norms) over time (Pollitt, 2001; Shankoff & Marshall, 2000). Children in the motivating and facilitated environment are more competent than children in a less stimulating and less rewarding environment (Fernald, Kariger, Engle, & Raikes, 2009). It has also been found that children with responsive caregivers and those who are in more stimulating environments are better advanced cognitively than children in less stimulating homes; parents who interact frequently with their children promote their cognitive, social, and emotional development (Shankoff & Marshall, 2000). These show that all items might not equally measure all students in various cultural, social, and biological contexts. In this regard, international tests have few items matching the ability level of the average low-income nation students (since these students are two standard deviations below the international mean, and most items are made to match the mean).

Many sub-Saharan African countries are shifting their junior primary science and mathematics education system towards more integrated science approaches by including skills such as scientific reasoning and solving real-life problems to cope up with rapidly changing and complex societies that demand critical decisions and judgments (Verspoor, 2008). The design of the curriculum frameworks of such countries tends theoretically in a way that classroom instruction and assessment procedures should follow constructive learning theory which promotes students' higher level of learning. This shift of educational system seems to match the students' learning with international trends. However, according to Joshi and Verspoor (2013) and (Verspoor, 2008), the actual classroom learning is characterized by low-level understanding, and classroom assessments and examinations of such countries are dominated with tests poor in quality and lack the quality to measure higher-order and problem-solving abilities. Even though the modern curriculum of such countries emphasizes higher skills of students' learning, current assessment practices are heavily limited to the recalling of memorized facts, re-stating, and definitions. Verspoor (2008) contends that such an assessment system leads students for rote learning, memorization, and recommends new assessment instruments to measure higher skills of students' learning, 21st-century skills, and newly introduced literacy. This suggests that there is a need for new assessment tools to explore students' way of learning, promote higher learning, and identify levels of reasoning which guides possible measurements towards students' higher-level learning based on the results of students' ability.

2.9 Assessments in the Ethiopian context

The Ethiopian government had launched an education and training policy that demands a focus on problem-solving and student-centered teaching to improve the quality of education and tackle the outdated teacher-centered approach (MoE, 1994). The ministry of education stated the main problems of education in Ethiopia are limited and inequitable access, lack of quality and relevance, and continuous decline in quality and standard (MoE, 2003). One of the causes for the decline in quality and standard of education was the teacher-centered approach (MoE, 2009). That is one of the reasons why the education and training policy of Ethiopia underlined the need for shifting from a teacher-centered to a student-centered approach (MoE, 1994).

In this regard, the Ethiopian curriculum framework designed in a way to encourage learning through dialogue that provides opportunities for classroom discussions in which students can exchange ideas, resolve misunderstandings in a variety of activities that allow them to construct meaning for themselves out of the information they are receiving (MoE, 2009). The curriculum framework adopts the principles of active learning and a competency-based approach to education. According to the framework, active learning involves doing, observing, and dialogue. The major goal of teaching physics according to Ethiopian grade 7 and 8 syllabi is, to help students think of physics not as an established body of knowledge, but rather as an active process of inquiry in which they can participate. It states that the learning and teaching process should be learner-centered, interactive, and participatory, and also the teacher should relate the learning experiences to the students' daily life. This implies that the framework encourages scientific reasoning and argumentation.

To materialize this, various programs were developed. Higher Diploma Programs (HDP) for teacher educators have been designed and provided in order to realize the implementation of student-centered approach (MoE, 2003). Furthermore, different staff development training such as Continuous Professional Development (CPD), Teachers' Development Program (TDP), and School Improvement programs (SIP) has been provided for School teachers (MoE, 2005).

This can, however, be best achieved when teachers use learner-centered, interactive, participatory and try to relate the content with day to day activities along with the proper assessment that can help them gather real-time information about students' learning. Proper implementation of active learning along with suitable assessments in a classroom situation is expected to produce students with the responsibility to construct high-level understanding, critical thinking, and a high ability of reasoning (Prince, 2004). To assess such students' abilities, one needs to employ construct-based assessment models that consider the investigation of both theory and classroom context to actualize the expected learning outcome.

The Ethiopian education and training policy underscored the continuous assessment in academic and practical subjects to ascertain the formation of an all-round profile of students at all levels (MoE, 1994). In this regard, various bodies have tried to prepare assessments and materials that help to develop classroom assessments. According to Arega (2014), assessment guidelines, techniques, and resource materials for the development of summative and formative assessments were developed by the organizations such as the National Organization for Examination (NOE), continuous professional development (CPD), and the Institute of Curriculum Development and Research (ICDR). However, the resources materials, and assessment guidelines were mainly focused on the traditional summative aspect of classroom assessment and heavily focus on the content and criteria. National educational assessment and examination agency (NEAEA) has prepared a comprehensive and system-wide classroom assessment manual that may enable school teachers to implement the competency-based curriculum in the classroom and improve their assessment techniques. NEAEA has been conducting regional exams at the end of grade eight to promote students into junior secondary school in Ethiopia. Early Grade Reading Assessment (EGRA) is also another assessment which has been developed in collaboration with USAID to obtain information about reading skill among grade 2 and 3 Ethiopian students. National Learning Assessment (NLA) is also one of the commonly known assessments which are conducted every four years to obtain information about the overall learning achievement of grade four and eight students. The assessments, however, focus on assessing performance skills, in general. It has no room to assess the skills of the 21st century such as problem-solving, critical thinking, and scientific reasoning (MoE, 2019). The results of such assessments are mainly aimed to pass or fail students from grade to grade or higher level of schooling instead of using the result to explore students' level of understanding, the pattern of reasoning, and for the improvement of students' learning. This suggests that the assessment system in the Ethiopian context does not agree with the stated goals of curriculum frameworks. Therefore, one can argue that the currently available assessments in the Ethiopian context are not devised by establishing an assessment framework, conceptualizing constructs to be tested, and considering students' learning progression.

The argument of this study, therefore, is the need for suitable assessments of students' ability of scientific reasoning and assessment items which is context-based, construct-based, domain-specific, and higher-order level assessments to explore and measure students' level of reasoning and the pattern of reasoning in order transform science education.

2.10 Assessment development and validation procedures

To develop and validate assessment tools, appropriate assessment development procedures that align with the construct to be measured need to be devised. The most commonly known assessments are item-driven and construct driven (Brown & Wilson, 2011). Item driven assessments contrasts construct driven assessments, in which, they tend to omit the construct definition and focus more directly on available items.

Most previous assessment procedures and assessment tools are criticized for focusing merely on item creation and validation neglecting the learners' cognition aspect, which is considered as a cornerstone of assessments, before item development (Brown & Wilson, 2011). They are likely to follow a path that consists of item development – scoring - analysis – scale development. Such assessments lack consideration of construct definition, what is being measured. For instance, the test development procedure by Oriondo and Dallo-Antonio (1998) that consists of the design of the test, trial test, determination of validity, and determination of reliability can be categorized into item-driven assessments. Such test development procedures, in the development and validation stage, ignore the conceptualization of a construct which is expected to be the main aspect of assessments. In this study, Wilson's (2005) assessment cycle and Senocak's Framework were used because they allow construct-driven assessment development and validation.

2.10.1 Wilson assessment cycle

Currently, it is suggested that the assessment of students' learning should be based on a theoretical conception of learner (Corcoran et al., 2009; Kind, 2013; Wilson, 2005). It is required first to establish a students' theoretical ability this, in turn, guides assessment development. There has to be a cyclical or iterative process that can help to know who learners are and what measures should be taken with the help of items that make trigger students' responses. It is better to decide after indentifying students' patterns of cognition than a one-shot approach (Wilson, 2009). One needs to use items to gather students' responses that can help to know the status of learner progress than to collect a final result for making a judgment of student competency after instruction has already been delivered. Instead of performing analysis on a test as a whole at a single spot, it is better to devise a way that can provide measuring the quality of questions to see how appropriate they are for the respondents and how well they measure their abilities. This leads to "construct-driven", "evidence-based" and developmental-perspective. Wilson's assessment cycle focuses on formative assessment instead of evaluating what students have performed after the instruction. It is helpful to follow students' progress and to identify what they can do and can't do at each level by identifying misconceptions.

The assessment development process starts with a definition of the construct. A construct is a theoretical model of a person's cognition such as their understanding of a certain set of concepts (Wilson, 2005). The construct map is a well-thought-out ordering of qualitatively different levels of performance in a domain (Wilson, 2005). According to Alonzo and Steedle (2009), different levels of students' abilities can be identified in a construct map by using standard documents and cognitive research. The levels can be derived from scientific findings, research results, students' alternative conceptions, and stated objectives in standard documents. The top-level represents currently accepted scientific understanding. Lower levels of the learning progression can be defined through a review of the research literature on students' understanding of the identified construct, including "misconceptions," (Alonzo & Steedle, 2009). Item design shows how the theoretical construct can be measured. According to Wilson (2005) items represent constructs that are to be manifested themselves in the real world. By using items, students' capabilities about constructs can be observed. Items need to be written in a way that can address all the qualitative levels of the construct map. Outcome Space shows how students respond to the items. This block describes the qualitatively different levels of responses of students to the items. This outcome space helps one to identify the students' responses in a construct map. This provides a visible representation of each theoretical level defined in the construct map. It is based on this representation that we can infer about the construct map. According to Wilson (2005), this outcome space is considered as a scoring guide. The measurement model describes a way to relate the scored outcomes from the item design and the outcome space back to the construct map. It relates individual scores and assessment items to the original construct map. It provides an understanding of theory (construct) and observation (data), and the relationship between these two.

Figure 2.2 represents the four blocks of Wilson's (2005) construct-driven assessment cycle to guide the process of test development and validation. It is based on modern test theory and starts with defining a targeted construct, and based on that assumption deciding what items should be used, then scoring criteria, and finally producing an assessment scale. However, this assessment cycle lacks explicit presentation of conceptualization of a construct being assessed with an extensive review of the literature.

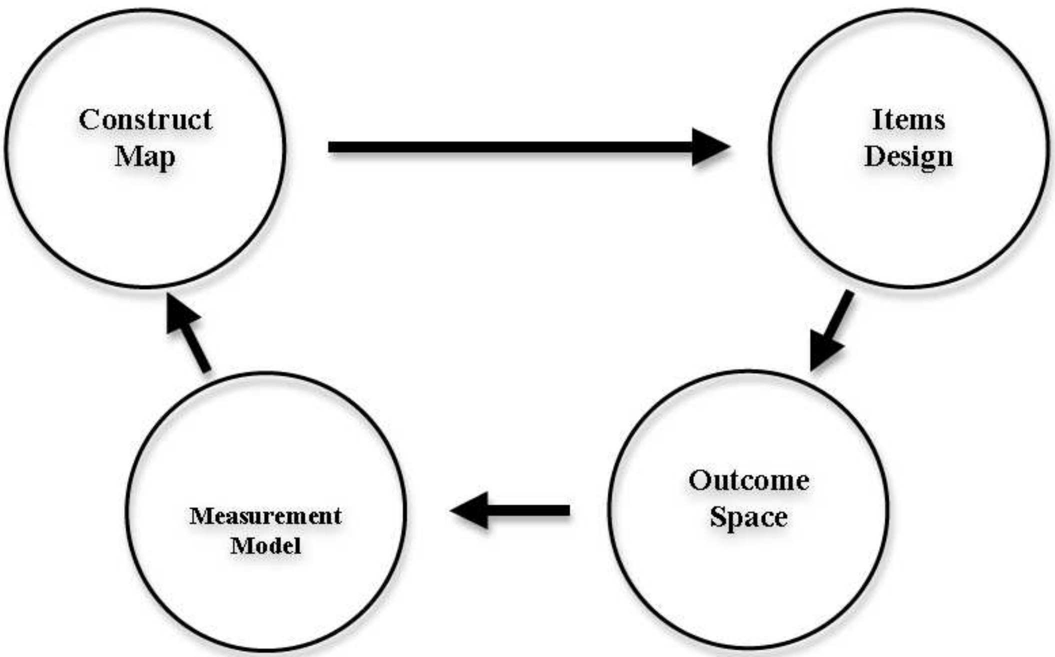


Figure 2.2: Construct modelling (Wilson, 2005)

2.10.2 Senocak's Framework

Among different construct-driven assessment frameworks, Senocak's framework is one of them. This framework explicitly presents each stage of item development by describing what to do in each step. The framework involves four stages of instrument development procedure: item formation, content validation, construct validation, and reliability calculation.

2.10.2.1 Item Formulation

This stage includes three steps: an extensive review of the literature (conceptualization of the construct to be studied), item development, and item modifications (Martin & Jamieson-Proctor, 2019; Senocak, 2009). The item formation stage involves an extensive review of the literature to conceptualize the construct to be measured in a way that represents students' ability, item development in a way to represent the conceptualized construct, and item modifications before administering to the selected sample.

2.10.2.2 Conceptualization of scientific reasoning

An extensive review of literature allows identifying what entails to the construct to be studied. Before the creation of items, the construct understudy needs to be defined explicitly. For this study, the construct to be studied i.e. scientific reasoning was defined in terms of learning progression and Styles Scientific of Reasoning (SSR). Scientific reasoning is conceptualized and four progressive levels were identified.

All aspects of human cognition are expressed by processes such as perception, attention, memory, categorization, logical reasoning, and associative learning considered as domain-general (Gelman & Noles, 2011).

Consequently, previous studies on scientific reasoning give more weight to the view that scientific reasoning is domain-general (Kuhn, 2002; Zeineddin & Abd-El-Khalick, 2010). Based on this assumption scientific reasoning is considered independent of content knowledge and it is assumed to be merely a skill to be developed. This view imposes the researchers to develop assessments that can measure students' ability of scientific reasoning skills such as control of variables, hypotheses generations, evidence generation, evidence evaluation, and drawing conclusions (Opitz et al., 2017) alone without emphasizing the knowledge. Such assessments are demanding to measure students' reasoning based on the reflections of the thinking processes instead of understanding scientific reasoning directly from the science culture where the reasoning has originated by considering various forms of knowledge. However, recently this view has substantially been challenged by the view that scientific reasoning is knowledge-dependent domain-specific (Kind & Osborne, 2017; Osborne, 2013; Zeineddin & Abd-El-Khalick, 2010). It is important to assess not only competencies of reasoning skills but also the styles of reasoning how such skills can be made manifest in real-life situations. To this end, recent studies give attention to relating styles of reasoning with the competencies of reasoning as well as all the forms of scientific knowledge instead of dealing with competencies of reasoning alone (Kind & Osborne, 2017; Osborne et al., 2018). They argue that giving much focus on reasoning skills alone could not help students to develop sustainable development in the whole picture of students' progress. Styles of reasoning are advantageous over the other assessments of scientific reasoning frameworks in that it recognizes the need for all three elements of domain-specific knowledge: content, procedural, and epistemic. Styles of reasoning are developed based on human thinking progresses and the development of natural science over more than two thousand years (Hacking, 1992). Even though thinking skills are transferrable, the ability to think and reason critically requires in-depth knowledge and understanding of a particular domain and is dependent on domain-specific content knowledge (Davies, 2013; Tiruneh et al., 2017). Scientific reasoning, therefore, in this study is considered as a subject knowledge-dependent, domain-specific, and expressed through learning progression and styles of reasoning. Scientific reasoning here defined by combining Ford and Wargo (2012) with McNeill and Krajcik (2011) frameworks in a way to contain the three types of scientific knowledge (level 1 and 2 – content knowledge, level 3 procedural knowledge, and level 4 epistemic knowledge) for the three styles of scientific reasoning.

2.10.2.3 Item development

After the conceptualization of the construct to be studied, items that represent all the levels of cognition need to be developed. Researchers who are engaged in measuring a construct have two alternatives. It is possible to adapt already existing and validated standardized tests in a way that can properly represent all the areas of the construct with modifications or one can develop new items in a way that can align with the levels of the construct. However, such standardized tests are criticized for they have not been conducive to measure skills of 21st-century lack to represent contextual factors (Goldman & Pellegrino, 2015). This indicates that it is better to develop new items by exploring real situations (contextual factors) such as relating learning with everyday practices and students' level of reasoning, and participating individuals who have direct or indirect relationships with the result of the assessments than using existing assessments. Recent studies recommend involving students, teachers, professionals in related areas, and experts in test development rather than participating individuals who have little knowledge about students' learning and ability (Brown & Wilson, 2011).

2.10.2.4 Item modifications

Before items are administered to the selected sample, it is required to revise the items by a panel of experts. Items can be modified firstly in terms of construct level such as a clear representation of the construct, alignment with the expected levels, and level of difficulty. This can be done by considering theoretical assumptions, life situation, and expected outcomes. Secondly, items can be modified in terms of distracter level in which whether the distracters are developed in a way that can represent various alternative conceptions and progressive understanding. Thirdly, item modifications can also need to consider the usage of language, clarity of writings, and length of statements.

2.10.2.5 Content validation

In the test development, one has to consider the content area from which a test is developed. It can be determined by the answer to the question, ‘‘Does the test contain items from the desired content domain?’’ This can be achieved based on the assessment by experts in that content domain. According to Oh, Seo, and Kozub (2010), face and content validity can be ensured using a panel of experts by evaluating items in terms of the contents to be covered.

2.10.2.6 Construct validation

Traditionally, validity refers to the extent to which the items measure what they are supposed to measure. In modern conception, one of the major types of validity is construct validity, which is the interpretations and uses of test scores that are validated, and not the tests themselves (Kane, 2006; Messick, 1998). Therefore, in the modern concept, validity is supposed to be established when evidence and theory support the interpretation of test scores (American Educational Research Association (AERA), American Psychological Association (APA), and National Council for Measurement in Education (NCME). The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretation. Construct validity is also considered as the most important type to be established and the base for other types of validity (Mislevy, 2007).

According to Burton and Mazerolle (2011), construct validity is the extent to which operational measure (the responses for the items) correlates with the theoretically conceptualized construct. In construct-based assessments, it is highly required to ensure construct validity because the focus of such assessments is measuring students’ ability in terms of conceptualized constructs. Therefore, to assess students’ abilities concerning a specific cognition, one has to use the proper item analysis model which can help to ensure construct validity. One way of ensuring construct validity is by employing the Rasch analysis model (Bond & Fox, 2015; Cavanagh & Sparrow, 2010a; Martin & Jamieson-Proctor, 2019).

2.10.2.7 Test specification and blue print

The testing process is expected to contain at least three phases: pre-testing phase, testing-phase, and post-testing (Al Fallay, 2018). Test specifications and blueprints are categorized under the pre-testing phase. Specification of test blueprint is an important issue in test development to decide the appropriate representation of content, difficulty level, and to have clear direction before the item creation. Test specifications and blueprints are part of construct validity. Test specifications provide guidelines for the content to be included in the test development and the way how items to be written; blueprint involves the design of the test and the number of questions that represent the construct being measured (Champagne, 2015).

In this study test specification and blueprint were prepared by referring grade 7 and 8 physics subjects objectives, syllabi documents, and literature in science teaching and learning along with subject professionals to determine the contents to be included, the students' levels of reasoning, the number of items to represent the levels of reasoning and design of the test.

2.10.2.8 Reliability calculation

In classical test theory (CTT) reliability is considered as consistency or precision of a test. For such a case, reliability is calculated using correlation coefficients such as Pearson correlations (for interval/ ratio level data) and Spearman correlations (ordinal level data). Unlike CTT, modern test theory such as Rasch analysis produces both item reliability and person reliability along with item separation and person separation. Person reliability along with person separation is used to determine whether the numbers of items are enough to separate the students into different levels of ability; item separation along with item reliability is used to determine the item difficulty in terms of expected levels of a construct (Boone, Staver, & Yale, 2013).

This study followed the combination of Wilson and Senocak's frameworks. The two frameworks were selected for this study because they allow a construct-driven assessment development procedure by first considering the conceptualization of a construct to be studied before item creation. However, Senocak's framework doesn't allow one to develop a construct map and items separately instead it presents both in the same category. On the other hand, Wilson's assessment framework doesn't provide validation procedures explicitly. For this reason, in this study, the two blocks of the Wilson assessment cycle (construct map and item design) were used to develop construct map and item development whereas the conceptualization of the construct and validation procedure was guided by Senocak's framework.

2.11 Classical test theory (CTT) vs. modern test theory (IRT)

One of the longstanding educational problems is developing good tests and questionnaires for educational assessors and researchers. Classical test theory (CTT) is established in the early 20th century (Traub, 1997). The advantage of CTT is the relatively weak theoretical assumptions that make it easy to apply in many testing situations (Hambleton & Jones, 1993). That is, it is simple to apply and it has few assumptions. It allowed quantitative analysis of reliability, but not sufficient to overcome issues like "guessing" and "careless" mistakes. The CTT also made an assumption that all test items make the same contribution to the overall score independently of their individual difficulties. That is the person statistic (i.e., observed score) is (item) sample dependent, and the item statistics (i.e., item difficulty and item discrimination) are (examinee) sample dependent (Wilson, 2005). Students' ability depends on item difficulty and vice versa. A student who scores 70% on a particular test may score less than or greater than 70% on other tests. A score value of 70% is difficult to deduce about one's ability. Modern tests can solve such problems by estimating difficulty level and ability level on the same scale (Wilson, 2005). Modern test theory deals with the ability of the students and the difficulty of the items as independent parameters, unlike CTT which has no way to separately identify these two constructs. This compelled some researchers to think and find out new and better statistical tools for validating tests and questionnaires (Lord, 1980; Rasch, 1960). One of the most serious problems in assessment is not only using statistical methods for analyzing items but also simply that test development has been far too item-driven (Corcoran et al., 2009). In education, it is particularly important that tests support good teaching and learning, and that tests assess what educators think is important (Kind, 2013). Assessors, therefore, also have been striving to find better procedures for "construct-driven" and "evidence-based" assessment and to developing assessments that support learning progression (Kind, 2013; Wilson, 2009).

All these issues are related to the modern test theory and resulted in indicators for good tests with a much broader perspective than CTT's concept of reliability.

Item Response Theory (IRT) provides three models: the one-parameter logistic model (1-PLM), two parameters logistic model (2-PLM), and three parameters logistic model (3-PLM) (Lord, 1980). The one-parameter logistic is called the Rasch model and the two are called IRT models. Although IRT and Rasch are similar to each other in terms of computation, they differ in approach to data/model fit representation. IRT and CTT are exploratory and descriptive models; Rasch's model is confirmatory and predictive (Bond & Fox, 2015). IRT is said to be descriptive because it aims to fit the model to the data. On the other hand, Rasch is prescriptive for it is required to fit the data into the model. Exploratory models must account for all the data, all data are required to be analyzed, interpreted, and explained. The IRT models are akin to statistical analyses in that additional parameters are added to the model to account for/explain the variability in the data rather than adhering to the properties of fundamental measurement and then questioning the data points that do not conform to those prescriptions. A confirmatory model requires the data to fit that model. Data that fail to fit the model are not required to be explained; they are only subject to modification or rejection. According to Bond and Fox (2015), the Rasch model is required to answer the questions: How well do the empirical data fit the measurement model requirements? Does the instrument yield invariant, interval-level measures for its intended purposes? For the Rasch model, we need to find a set of items that satisfies the model while for an IRT model we need to find a model that fits an existing item set. This shows that it is easier to find items that fit a model by iteration process than finding out the model that fits an existing set of items when we want to develop items that help to measure a certain construct. Moreover, the Rasch model has an advantage over IRT in that it can put the item and examinee attributes on the same scale and the presence of sufficient statistics that can help for Rasch estimates.

2.11.1 Rasch model

Rasch measurement theory was developed by George Rasch to design and revise assessment instruments (Bond & Fox, 2015). The theory allows one to determine the validity and reliability of standardized tests. In education developing an instrument that measures a single trait is a challenging task. However, the Rasch measurement model can help one to develop a single variable by evaluating the quality of items in relation to a defined variable. Rasch's analysis assumes that the items involved in measuring different aspects of a single trait. It provides indicators that can allow determining items that can contribute to measuring the predicted construct and items that do not fit to represent the predicted construct. In this study, Rasch modeling techniques were used to ensure construct validity by evaluating the fit of the instruments' items to the underlying construct. Rasch analysis provides indicators that allow evaluating the quality of items whether they represent a construct, gives the relationship between construct theory, a data, and tells us whether the items measure a single dimension of intended construct theory (Baghaei, 2008). This allowed the researchers to identify items that fit the model, items that need revisions and items to be discarded from the tool.

In this study, Rasch one parametric model was used because Rasch analysis provides indicators that help to develop construct-based assessments and ensures construct validity (Bond & Fox, 2007; Cavanagh & Sparrow, 2010b; Martin & Jamieson-Proctor, 2019).

2.12 Validity argument

The educational measurement uses limited samples of observations to draw general and abstract conclusions about persons or other units (Kane, 2006). To validate an interpretation or use of measurements according to Kane is to evaluate the rationale or argument, for the claims being made, and this, in turn, requires a clear statement of the proposed interpretations and uses and a critical evaluation of these interpretations and uses. Ensuring that educational assessments have high validity is a fundamental aim of all those involved in the development of assessments (Shaw & Crisp, 2012). The essence of validity is to know that the inferences being made from results and the ways that results are used are appropriate. In the new definition of validity, there is more information about validity discussion and there are different ways of evaluating the quality of a test. Validity has changed from estimating the accuracy of the measure to the argument about its relevance. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretation.

Therefore, construct validity is said to be established when the items measure students' understanding of the selected constructs properly. To ensure this we need to:

- Use theory to describe the nature of the construct;
- Establish evidence that the test is measuring the construct.

We need to first establish a theory about that construct (what students' are expected to do, prediction of what students are supposed to do), and based on that theory we need to have evidence (what students achieved practically with the help of items and analyzing and interpreting the data). We can have three things to be considered: theory (expectation), evidence (observation), and procedure (the way we score, analyze. and interpret the data to relate theory and evidence). Construct validity is said to be established:

- If a test measure what it is supposed to measure (the relation between theory and test, does the test properly represent the students' understanding).
- If we can trust the interpretation made from the test (the procedure we follow is appropriate).

Traditionally validity concerns only what the test measures what is supposed to measure. It does not concern about the evidence of the interpretation based on the established theory. However, evidence-based validity deals with both what is supposed to measure and interpretation based on the established theory. To achieve this, theory (what students are supposed to do) should be established and data needs to be interpreted and also gives evidence to what extent is the interpretation evidence-based. That is, validity determines to what degree inferences based on test scores are appropriate and meaningful

(Kane, 2001), and to what extent is one's interpretation appropriate and meaningful. For SRPT the validity of the test was ensured based on two aspects of construct validity: Firstly, the fit of the test and unidimensionality was examined using Rasch analysis to determine whether SRPT measures the construct under study. Secondly, the distribution of the items in terms of theoretical assumptions was examined using an item-person map. These were further validated by expert reviews.

To sum up, this chapter has presented a relevant review of literature for the construct scientific reasoning, assessment of scientific reasoning, and procedures for the assessment development. It began by introducing the views of scientific reasoning, the ways how to understand scientific reasoning, and conceptualizing scientific reasoning. Then, theoretical frameworks were discussed and this yielded the conceptual framework that guided this study. Following this, the concept of assessment, global trends of assessments, and the nature of assessment were also discussed. Finally, the assessment procedures were discussed along with the assessment tools. In the following chapter, the method of how the assessment tool was developed to measure the construct scientifc reasoning will be discussed.

Chapter 3 Methods

This chapter, first describes the research perspective, the research method, the research design, and the research settings of this study. Then, discussions about population, sampling techniques, the construct development processes, and the overall SRPT development and validation will be followed. Besides, this chapter presents data collection procedures and analysis techniques used in this study. At the end of the chapter, validity and reliability issues will be briefly discussed.

3.1 Research perspective

In educational research settings, the philosophical worldviews of the researcher guide the selection of research design, the subsequent data collection, and analysis methods deployed in the study. Hence, it is important to communicate the philosophical perspectives of the researcher. In addition to this, this section also presents the ontological and epistemological assumptions held in this study.

Traditionally, there are two perspectives about reality. One of the groups believed that reality is subjective. To this group, reality is multiple ontological existences comprehended in a number of different ways by different observers. This group considers reality to exist in the knower's mind, thus requires meaning and interpretation directed toward certain objects or things, therefore, one needs to focus on the subjective point of view, experiences, and feelings (Creswell & Guetterman, 2019). In educational research, this paradigm is known as the constructivist worldview. The other group takes reality as a single ontological existence that exists 'out there' independent of the knower's mind. This group assumes reality as single, objective, and that follows unchanging laws. Hence, 'truth' can be known using the procedure of scientific methods. In education research, the objective reality camp is known as the post-positivist paradigm.

However, recently a third kind of paradigm, namely pragmatism, has become poupular. Pragmatism rejects the division between post-positivist and constructivist world views. In addition to this, pragmatism is not founded on the assumption of whether reality is solely a social construct as constructivists presume or ‘real’ as positivists believe.

The pragmatic view suggests that neither established theories nor meanings behind established knowledge independently represent one’s reasoning (Constantina & Constantine, 2014). Pragmatists believe that it is impossible to understand the reality about the world using either the objectivists view or the subjectivists view alone (Kivunja & Kuyini, 2017). Rather, pragmatists support a world view that considers the suitability and appropriateness of the selected approaches to solve the problem. This dissertation is guided by the pragmatists’ worldviews and hence exploit and incorporate the strong points of both the post-positivist and constructivist paradigm. The author of this study favors pragmatism because he shares the view that reality is found independent of the individual but it requires the individual’s interpretation and meaning-making about that external reality.

The current study attempted to construct, develop, and validate new instruments to assess grade eight students’ level of scientific reasoning. In line with this, what is the nature of reality about grade eight students' level of scientific reasoning? Do all students have a scientific reasoning pattern or do they have different levels of reasoning – naïve, mixed, and scientific – reasoning pattern? Different researchers approach such questions differently based on their philosophical orientation. For example, post positivists may hypothesize that grade eight students have a ‘scientific’ level of reasoning and then continues to develop and validate a test instrument that verifies or refutes this hypothesis. Whereas, constructivists do not assume anything beforehand. Rather, constructivists attempt to understand the layers of scientific reasoning observed during discussions with the research participants, and thus develop the instrument that provides evidence of multiple levels and patterns of students' scientific reasoning.

Human reasoning is relative and hence describes what goes on in a time-bound situation or context. Good reasoning is intertwined with valid arguments at that instant moment, though it can be changed, refuted, or improved afterward based on the evidence generated from recently developed scientific knowledge. The researcher believes that consideration of context, meaning, and interpretation along with the established theories are important to this study. Hence, this study argues for scientific reasoning test development that considers students' reasoning based on the levels of reasoning as reported in previous scientific theories and research findings along with contextual factors such as students' views of scientific knowledge, way of constructing knowledge, and students' views about the nature of science that can affect their reasoning. Hence, the pragmatic position was an appropriate philosophical stance to guide the study.

The pragmatist paradigm provides educational researchers to mix ontological and epistemological assumptions as well as the methodological processes of the research. This dissertation combines the constructivist qualitative methods and the post-positivist quantitative methods. This study was aimed to develop and validate SRPT which in turn used to investigate students' reasoning ability in terms of the four scientific reasoning levels conceptualized under the literature section (In particular see Section 2.2. 4). The aim of qualitative study was to explore students' ways of reasoning in relation to each level of reasoning, to decide what type of items needed for each level, and what types of alternative conceptions students hold. To explore students' pattern of reasoning for each level of reasoning students' were interviewed, discussions were held with primary school physics teachers, and contents to which items to be developed were identified. This qualitative study guided the development of items and chooses of item distracters. After exploring students' pattern of reasoning for each level, the items that represent each level were developed, revised, administered for the selected samples, analyzed, validated and students' abilities of reasoning were investigated.

3.2 Research design and method

One of the fundamental goals of science education in K-12 schools is to develop students' scientific reasoning skills. In this research, due emphasis is given to conceptualize the scientific reasoning construct, to develop corresponding science reasoning learning progressions and SRPT instrument, and to validate this test tool.

The current study focused to develop a research-based scale or test instrument that allows students to be categorically assigned along a continuum of scientific reasoning. To further elaborate the learning progression of the scientific reasoning construct and the test instrument developed, the following research questions were raised in Chapter One of this study.

The research questions were:

1) Which scientific reasoning skills are addressed by the construct map developed for the SRPT test items?
2) How are the content and construct validity of the SRPT items qualitatively established?
3) How effective are the SRPT test items in measuring grade eight students' scientific reasoning in physics?
4) What is the trend of Grade 8 students' progression among the different reasoning subskills measured by the SRPT?

Some of the research questions were related to identifying scientific reasoning skills addressed by a construct map which served to develop items and establishing content and construct validity of the items qualitatively, whereas others were focused on establishing the construct validity of SRPT qualitatively and quantitatively as well as in looking at what the data (students test scores of SRPT) suggest about grade eight students reasoning levels.

The development of SRPT items corresponding with the different levels of learning progression requires not only to construct conceptualization and item development but also detailed information about how grade eight students try to show reasoning of a scientific phenomenon using physics concepts. Combining both methods provides a broader perspective about the nature and character of students' scientific reasoning in physics. Moreover, triangulation of data collected is more likely to lead to the development of comprehensive research-based SRPT tools to answer the research questions and to greater validity of findings. In summary, the current study deployed both quantitative and qualitative methods to answer research questions regarding item functioning and construct validity in the instrument development process. In addition to this, the construct, i.e., scientific reasoning, conceptualization, item development, and instrument validation phases are guided by the pragmatic worldview.

In light of these underlying justifications, this dissertation opted for a mixed methods research design. The study uses, in particular, a sequential exploratory design firstly exploring students' patterns of reasoning using phenomenography (qualitatively) which is followed by investigating students' levels of reasoning using descriptive survey research (quantitatively). Thus, the research approach used in this study can be considered as an exploratory sequential design.

The main reason to blend these two methods was that qualitative data obtained in the first phase would provide useful context to the development of SRPT as well as to the quantitative data deployed in the second phase. The qualitative part elicits grade eight students' depictions of the basic nature and character of scientific reasoning and uses these descriptions to construct categories that form the different views of students' scientific reasoning pattern. Meticulous descriptions of students' scientific thinking and teachers' experience of their students' nature of reasoning has provided contextual aspects to the SRPT assessment instrument. In addition to this, students' ways of reasoning were thoroughly explored during instrument design and development in order to gain insights into students' scientific reasoning patterns about various physical concepts and phenomenon.

The use of this study utilized a mixed-method research design, which is consistent with the selected research paradigm, the nature of the investigated research problems, and the type of research questions. Thus, both qualitative and quantitative methods were deployed in the first and second phases of this study respectively. The first phase, which is the development of reasoning patterns and items, used a qualitative approach. The qualitative part aims mainly to gain insight into the nature and characteristics of students' scientific reasoning. For this purpose, the qualitative part chooses phenomenology because this approach enables researchers to reveal multiple ways that students reason about certain physical phenomena. Phenomenology serves as an analytical framework to understand students' patterns of scientific reasoning in relation to the four identified levels of reasoning. Hence, this study employed the phenomenological approach. The phenomenological approach used students' interviews and teachers' focus group discussion to organize the pattern of students' responses and teachers' views, opinions, and experiences about students' classroom physics reasoning. The interview enabled the researcher to amass detailed information from students about their level of reasoning. Moreover, the interview provided vital information about students' prior knowledge and also supported the test development to be contextual by highlighting the depth and breadth of content coverage in their physics lessons. The focus group discussion provided insights about teachers' beliefs, experiences, and views about their students' way of physics learning and reasoning. Furthermore, an in-depth analysis of students' interview and teachers' FGD were used to develop the construct map and the test items across the four scales and different reasoning levels.

In the first phase of this study phenomenographic approachwas used to explore patterns of responses and associated reasoning. Phenomenography is a research approach for examining different perspectives on reality; it refers to a method for describing the various ways in which a group of people perceives a phenomenon (Larsson & Holmström, 2007; Marton, 1981). Some considered phenomenographic study as a strategy of inquiry in which the researcher identifies the essence of human experiences, conceptualizations, perceptions, and understanding about a phenomenon as described by participants (Creswell, 2013; Marton, 1986). The goal of phenomenography, according to Marton (1981), is to investigate the differences in people's perceptions of a given phenomenon in the surrounding environment, and the investigation is based on the variation in people's approaches to understanding the phenomenon rather than the phenomenon itself.

One of the goal of this study was to investigate the students' views in relation to the four levels of scientific reasoning. Phenomenographic study was used to investigate the students' patterns of reasoning and underlined views. The use of phenomenography in this research is due to the insights of the phenomenographic approach that allows one to examine the variations of views participants hold about a phenomenon, an idea, and conceptualizations of an object of study (Ayene, Kriek, & Damtie, 2011; Marton, 1986). It is, therefore, an effective method to investigate the views of students against different levels of progression of scientific reasoning and to identify how the views of the students are linked to the different levels of scientific reasoning.

The second phase involves pilot testing, validation, and field-tests to assess students' reasoning abilities using SRPT. The validation procedure of the construct map and the tool related to tests required quantitative methods. To analyze the quality of assessment tools, quantitative approaches are often preferred. The quantitative part employed a descriptive survey research method. In the second phase, SRPT was pilot tested. Using statistical procedures the construct validity and reliability of the SRPT were explored and then field-tested. The descriptive research design was appropriate for the quantitative part of this study due to many reasons. The first reason is that this research design can characterize the study population thoroughly by going beyond the initial observation of the test scores. The second reason is descriptive study can lay foundations for subsequent research. Finally and most importantly, mixed methods research was appropriate in this study because the quantitative and qualitative methods suit for the purpose of development of construct map, tool design, development, and validation; and measuring the student's scientific reasoning. Therefore, in this study mixed approach was used. In this second data collection phase, the tool was administered by the survey research method. The survey method is an appropriate method to obtain information from large samples of the population at an instant time (not a time series or longitudinal). Survey methods require relatively minimal cost and relatively easy for making generalizations (Bell, 1997). Descriptive survey research was conducted for this study because it was a suitable method to collect data from a group of people on a given subject. The SRPT was administered for the selected sample and analyzed using Rasch analysis.

To sum up, educational research employed various mixed methods designs. Sequential and concurrent mixed methods are among the common ones. The selection of research design is determined by the type of problem to be dealt with and research questions to be answered. In this study, the researcher aimed to develop a new instrument that assess grade eight students' ability of scientific reasoning in physics and to assess students' level of reasoning. This study initially explored the reasoning pattern of students and develops tools.

It also explored teachers' views and experiences qualitatively, develop, and proceeded into the validation phase.

Sequential Exploratory Mixed design was a suitable method for such tasks that required the development and validation phases. This is because, in this study, the first phase involves exploring the phenomena with a few participants. By exploring the students' reasoning pattern and teachers' views about students' way of reasoning, the students' pattern of reasoning was explored, and also a tool that represents these were developed. In the first phase, qualitative data were collected and analyzed which were followed by the second phase of quantitative data collection and analysis. The qualitative findings using students' interviews and focus group discussions with teachers served as a guide for the development of students' reasoning and the tool that represents each reasoning level. Furthermore, the mixed-method enables the triangulation of quantitative and qualitative data to converge information (Creswell & Clark, 2018) and gain evidence about the validation of the instrument. It also helped to overcome the limitations of using either only quantitative or qualitative methods. Therefore, sequential exploratory mixed methods approach that mixed phenomenographic and survey was appropriate as a design for this study.

3.3 Research setting, population, and sampling

3.3.1 Research setting

This research was conducted at Hadiyya zone, Southern Nation Nationalities, and People Regional (SNNPR) government in Ethiopia. Hadiyya Zone is located in SNNPR whose administrative city, Hossaena is 234 kilometers from the capital, Addis Ababa. There were two reasons to select this research site. One, the researcher visited several schools at Hadiyya Zone while tutoring college students. From their responses to their teachers' questions, it was observed that students have reasoning related problems. This motivated the researcher to study the students' reasoning. Two, it was one of the sites of the TPSS project. TPSS is a research project that is engaged in improving the pedagogy of STEM subjects in Ethiopia. The researcher was one of the participants of the project. The research project has also found that students have poor reasoning ability and conceptual understanding. These two reasons motivated the researcher to study students' reasoning abilities and develop a tool that can measure students’ reasoning abilities which could be scaled to others since they use the same curricula.

3.3.2 Sources of data

Data for this study was gathered from primary and secondary sources. As primary sources subject experts, primary school teachers, and students were included. Experts checked the relevance of the content, the clarity of items, and the appropriateness and suitability of test items for the identified levels. They also provided necessary comments in revising the tool and the construct map. Teachers are the implementers of the contents of the subject. They are expected to know about the content, students’ levels, and students’ alternative conceptions and reasoning. They participated in the development of the tools and construct map. Students are the ones whose reasoning abilities and performance levels were assessed and the tool was administered for its validation. Physics curricula, textbooks, syllabi, lesson plans, and teacher guides for contents and classroom tests were used as secondary data sources. Tests used by teachers, national learning assessments, grade 8 regional exams, and physics text and reference books, TIMSS reasoning assessments were used as secondary sources of data.

3.3.3 Population

The population of this study are grade 8 students and teachers who teach physics subject in grades 7 and 8 at Hadiyya Zone. The selection of 8th-grade students is because grade 7 is an early stage of the partition of science into three independent subjects, the beginning of physics to be learned as a subject in the Ethiopian curriculum, and better maturity level of grade eight students than grade 7. In addition to this, National Learning Assessments (NLA) and regional grade 8 exams were also considered for the tool development for this study.

3.3.4 Sample and Sampling

To select the participating schools, teachers, and students, both purposive and random sampling was used. Six schools were selected by stratified random sampling which allowed the researcher to include schools in different locations randomly. The schools from town and rural areas were included in the sampling. Two hundred and three students from a school not included in the actual study participated in the pilot study and two hundred and forty-two students participated in the final study. To reach the recommended sample size for one parameter Item Response Theory (IRT) model which is 200 (Wright & Stone, 1979). One section per school was selected randomly for both the first and second-round testing.

Six primary school physics teachers were selected to conduct focus group discussions. (Creswell, 2012) recommended six to eight sample sizes for focus group discussions. The number of participating teachers was limited to six because this provides the researcher to manage the discussion and obtain necessary data from the participants. The teachers who were engaged in teaching grade eight physics subject. Moreover, the teachers took training on developing activities to teach in dialogical teaching methods, scientific reasoning and argumentation, and conducting scientific argumentation in the science classroom. The teachers have also implemented the aforementioned activities in classrooms for a year.

For the interview, 12 students were selected by purposive sampling where the purpose was previous achievement that helped members from varying achievement levels which includes high, medium, and low achieving students. Marton and Booth (1997) recommended sample size between ten to twelve people for phenomenographic study. The number of students was limited to 12 because the needed data was saturated (redundancy of information was recorded) by this number.

3.4 Data collection Instruments

For qualitative data gathering, students' interviews and focus group discussions (FGD) were used. For quantitative data gathering, scientific reasoning test was developed.

3.4.1 Interview

The interview was used in this study to explore grade eight students' scientific reasoning patterns in physics. Interviews served as effective tools for collecting in-depth data to understand the meanings of a construct under focus (Nunan, 1992). To collect qualitative data from the students, semi-structured interview questions were developed. Initially, five questions were developed by the researcher (See Appendix D).

The interview questions were developed to obtain students’ general view about scientific knowledge and the four scientific reasoning progresses levels: *factual knowledge, explanation of factual knowledge, evidence-based reasoning, and drawing conclusion*. The first interview question was designed to explore students’ general views about students’ scientific knowledge before and after learning physics concepts. The second interview question was developed to elicit students' views about scientific facts. The third interview question was developed to elicit students' information about how scientific explanations are generated and how it is related to daily experiences. The fourth interview question was developed to elicit students' view about scientific evidence and the last research question was developed to explore students' views about the scientific conclusion and how scientific conclusions is drawn. Under each question, students were asked to give examples from physics content knowledge which served to categorize students' patterns of reasoning for each level. Comments were received from two Ph.D. candidates in physics education at Addis Ababa University. The original questions were revised and one more question that can help to explore the probable trend of how students' reasoning ability grows was included in the final tool by the suggestion of one of the Ph.D. candidates. The interviews were conducted in either laboratory rooms or pedagogical centers of the primary schools in order to make students freely respond to the questions. The researcher interviewed by first requesting their consent and the interviews were audio-recorded. The interviews took between 10 to 20 minutes to complete the recordings.

3.4.2 Focus group discussion (FGD)

Focus group discussion is an appropriate tool to gather information from individuals who have similar backgrounds and experiences. Focus group discussion, which provides flexible and information-rich data, allows the emergence of a useful discussion about a broad topic (Dornyei, 2007). Focus group discussion was used, in this study, to collect data related to the belief, ideas, and experiences of teachers about students' reasoning progress when teaching physics. The questions for FGD were the same questions of the students' interview but modified in a way that can help to generate information from the teachers about students' views and patterns of reasoning (see Appendix E). In addition to this, the FGD with primary school teachers served to triangulate and corroborate the data obtained by students' interviews. With the help of FGD, the teachers' understandings about the students' levels of understanding, reasoning, and related alternative conceptions were identified.

The discussion was held in the physics laboratory room at Hossaena College of Teacher Education. The place was selected for its quietness which is appropriate for such a discussion. In the FGD, six teachers participated. All the participants reflected their views about students' reasoning with related examples from their experiences.

3.4.3 Objective items

To study students' levels of reasoning, this study has developed and validated the scientific reasoning tool for grade eight students in physics subject. The developed instrument was comprised of multiple-choice items which consist of one correct answer and three distracters. The items were designed in a way that can represent students' reasoning ability at each level. The correct answers represented the currently accepted scientific knowledge, whereas the distracters were developed based on the qualitative findings and alternative views identified in the literature. The item development procedures are discussed in the following sections.

3.5 Assessment development procedure

In this section, elaborated discussions regarding the design, the collection of data, and the analysis of the SRPT are presented below. As discussed above, this study combined Wilson's assessment cycle and Senocak's frameworks for the development of the SRPT (see Section 2.10 for details). The assessment development procedure involves conceptualization of the scientific reasoning progresses, development of items, content validity, construct validity, and judgmental analysis.

Figure 3.1 represents the assessment development procedure constructed by considering Wilson's assessment cycle and Senoack's assessment framework. The conceptual framework in Figure 3.1 served as guidance for the development of the construct map by conceptualizing scientific reasoning as discussed in chapter 2. To develop the initial construct map literature was reviewed and four levels of scientific reasoning progress, indicated in Table 2.5, were identified along with the expected performance for each level. The Construct map, in turn, guided the development of items that represent each level of students' reasoning progress.

The researcher provided training for the teachers on what a construct map is and items that represent the levels. Having this, the researcher dispatched the initial items developed by the researcher and a construct map to the teachers to come up with new items and modified items. The teachers also introduced Haladyna, Downing, and Rodriguez's (2002) multiple-choice item-writing guidelines. Haladyna et al. (2002) consist of 31 multiple-choice item-writing guidelines. Based on the guidelines, the discussion was held on how to construct multiple-choice items, stems, distracters, and correct answers vis-à-vis students' ability levels. The teachers came after three weeks with new items and modified items. The discussion was made and items were modified, rejected and new items were incorporated into the tool. In addition to this, the test items and construct map had passed content validity, construct validity, judgmental stage, and field testing before the final test and construct map comes out. The detail of each element of the assessment procedure depicted in Figure 3.1 discussed in the sections below.

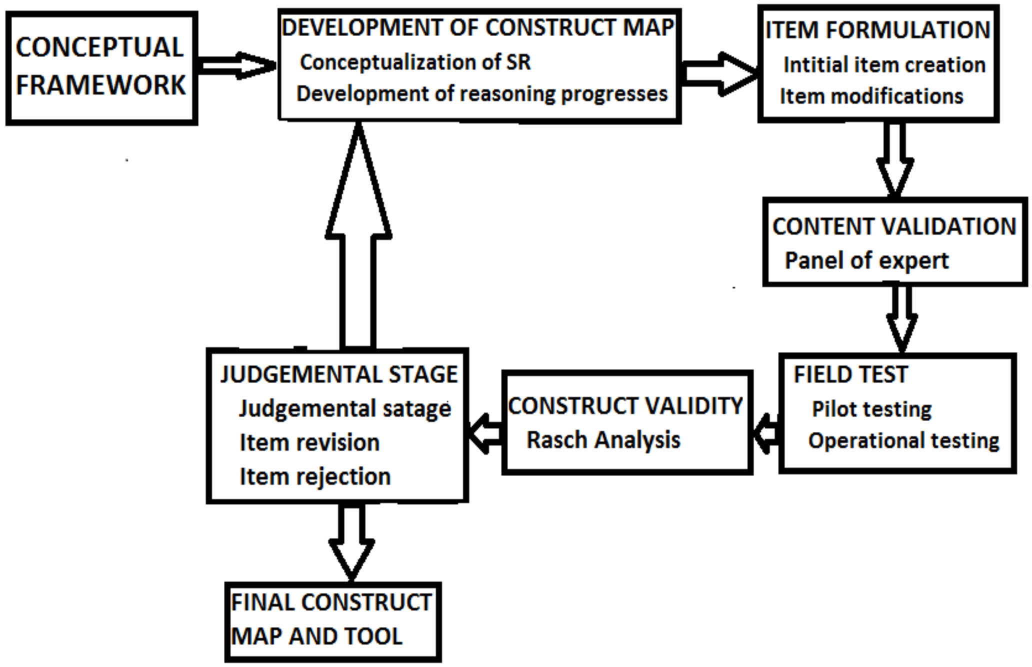


Figure 3.1 Test development procedure for test SRPT

3.5.1 Development of the construct map

The person-item map demonstrated in Figure 4.1 was developed by considering the conceptualization of scientific reasoning, students' interviews, and FGD with primary school physics teachers (see Section 4.1.3). The person-item map comprises four levels of students' reasoning: factual knowledge, explanation of factual knowledge, evidence-based reasoning, and drawing conclusion (Table 4.10) along with the students' views (see Table 4.12) towards each level of scientific reasoning (see Section 4.1.3 for the detail of how person-item map was developed).

3.5.2 Item formulation

This stage included three steps (see Section 2.10): an extensive review of literature on the conceptualization of the construct to be studied (see Section 2.2.4), item development, and item modifications (Martin & Jamieson-Proctor, 2019; Senocak, 2009).

Items were developed from the content of grade 7 and 8 physics subjects from which students learned to represent the levels in Table 2.5. Initially, the researchers developed 44 items by referring to the National Learning Assessments (NLA) of Ethiopia, TIMSS reasoning items, grade eight physics classroom tests, physics grade 7 & 8 textbooks, and a reference book (Hewitt, 2006). The items were developed from the content students learned in grades 7 and 8. The item development followed Haladyna et al.'s (2002) multiple-choice item-writing guidelines. Each item consists of three distracters and one scientifically correct answer. The distracters were developed to include correct responses, partially correct responses, incorrect responses, and naïve responses (see Section 4.1.3 for the detail of item development).

Initially, eleven questions (each consists of four items for each level) were developed; a total of 44 items were developed. The items were written in a way to contain one parent question which consists of four items to represent the four levels of reasoning (see sample items). The four items are related to one another. For example, there is question 1; under question 1 there is an item 1.1 to represent level one, item 1.2 to represent level two, item 1.3 to represent level three, and item 1.4 to represent level four. This approach forces students to establish the relationship between each level to answer the four items under each question. Such way of construction of tests requests students to develop the relationship between the factual knowledge (at the first level), the explanation for the factual knowledge (at the second level), generation of evidence for the explanation (at the third level), and drawing conclusion based on the evidence (at the fourth level) instead of picking a correct response in the usual multiple-choice questions. This type of item development was used in this study because (1) it allows students to think and reason in creating a link between the four levels instead of merely selecting a single answer (2) to have coherence and flow of levels consistently based on a single concept instead of writing different items from different concepts. This enabled the researchers to achieve one of the arguments of this study which contends that scientific reasoning is dependent on the three types of knowledge and needs to progress from the lower level to higher for a particular concept instead of viewing it as a mere skill independent of content knowledge. The first and the second level were constructed to represent content knowledge; level three is was intended to represent procedural knowledge, and the fourth level was to represent epistemic knowledge. The researchers followed Haladyna et al.'s (2002) multiple-choice item-writing guidelines to write SRPT multiple-choice items. All the questions include three distracters and one scientifically correct answer. The distracters were developed in a way to include a dominantly naïve view, mixed but towards the naive view, mixed but towards the scientific view, and dominantly scientific view.

3.5.3 Sample items

Item 8 presented below is meant to measure students' ability in data-based reasoning. At the first level, item 8.1 is designed to assess the factual knowledge. At this level the students are expected to identify the units of measurements, in this case, to identify the unit of voltage (volt). At the second level item 8.2 is designed for the students to provide explanation about the relationship between variables, in this case, to explain what the students performing the experiment want to know, that is, the relationship between voltage and current. At the third level, item 8.3 is designed for the students to provide evidence as to why they provided the explanation for the second level by using the data. In this regard, the students are expected to generate evidence from the data (the measurement is taken on voltage and current at constant resistance) to know the relationship between voltage and current. Finally, at the fourth level, item 8.4 is designed for the students to draw conclusion by observing the pattern of data given in the table. If the students are able to give the relationship between the voltage and current by the given values (1.5V, 3V, 4.5V, and 6V), they are expected to give their relationship for other values of the voltage (in this case 7.5V).

Question 8: Two students perform an experiment. To accomplish this, they collect different working batteries, wires, and ammeter. After connecting properly they measured the values of current (I) in the circuit by varying the number of batteries for a single value of resistance (R). The students' measurement is tabulated as shown below:

Voltage (v)	Resistance (Ω)	Current measured by first student (A)	Current measured by second student (A)
1.5	2	0.73	0.76
3	2	1.40	1.45
4.5	2	2.20	2.20
6	2	2.95	3.05

8.1 Which of the following unit of measurement is used to measure voltage? (Factual level)

a) <u>volt</u> b) ampere c) ohm d) ammeter

In the above experiment what do the students want to know? (Explanation level)

a) <u>They want to know the relationship between voltage and current</u>
b) They want to know how resistance affects current.
c) They want to know the relationship between resistance and current
d) They want to know how resistance affect voltage

8.3 Why do you select the above answer for question number 8.2? It is because (Evidence level)

a) <u>The measurement is taken on voltage and current at constant resistance</u>
b) The measurement is taken on current by varying voltage at constant resistance
c) The measurement is taken on voltage and resistance is calculated

d) The measurement is taken on voltage, current, and resistance.

8.4 What would you expect best if the students use a 7.5 volt battery in the above experiment? (Drawing conclusion)

a) The value of the current measured by the students will vary in the range between 2.95 and 3.05A.
b) The value of current measured by the students will be less than 3.05 A
c) The value of current measured by the students will be less than 2.95A
d) <u>The measured value of current will be greater than 2.95A and 3.05A</u>

Item 9 is meant to measure students' ability in mathematical reasoning. In the first level (9.1) it is designed to assess the factual knowledge (identifying the given variables), in the second level (9.2) students are required to reason mathematical variable relationships, and in the third level (9.3) they are required to provide evidence as to why they provided the explanation for the second level. Finally in the fourth level (9.4) students are expected to draw a conclusion.

Question 9. A man pushes a body with a force of 20N and produces an acceleration of $2m/s^2$.

9.1What are the given variables in question 9?

a) Force and velocity **<u>b) Force and acceleration</u>**

c) Mass and acceleration d) Force and speed

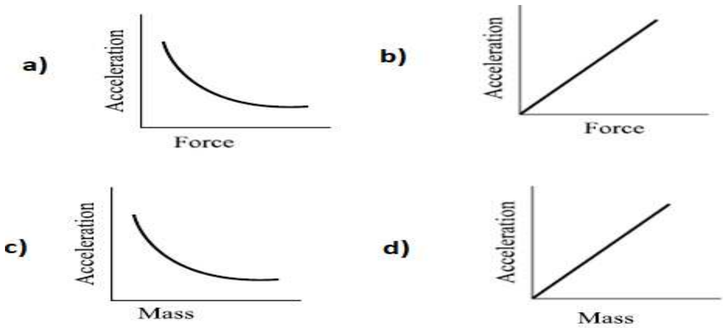
9.2 What force will produce an acceleration of $4m/s^2$ for constant mass in question 9?

10N b) 20N **<u>c) 40N</u>** d) 80N

9.3 Why do you select the above answer? It is because

a) <u>There is direct relationship between force and acceleration</u>
b) There is inverse relationship between mass and acceleration
c) To determine its acceleration mass should be given
d) The same force can produce different acceleration whether mass changes or is constant

9.4 Which one of the following graphs represents your answer to question number 9.3?



Item 5 is meant to measure students' ability in model-based reasoning. In the first level (5.1) it is designed to assess the factual understanding, in the second level (5.2) students are required to provide an explanation about the relationship between variables, and in the third level (5.3) they are required to provide evidence as to why they provided the explanation for the second level by using the data. Finally in the fourth level (5.4) students are expected to draw a conclusion based on the explanation in the second level and evidence in the third level.

Question5: Suppose an apple hung on a tree.

5.1 What type of energy does the apple hold at the highest point?

a) Potential energy b) Internal energy

c) Kinetic energy d) Elastic potential energy

5.2 If the apple falls down, what will happen to the energy of the apple just before the apple hits the ground?

a) Kinetic energy will be converted to potential energy
b) The type of energy at the top will remain the same
c) Potential energy will be converted to kinetic energy
d) Some amount of energy will be destroyed

5.3 Why do you select the above answer for question 5.2? It is because

a) As the apple moves down its height increases but speed decreases
b) As the apple moves down its height decreases but speed increases
c) As the apple moves down both its height and speed decreases
d) As the apple moves down its height decreases but speed remain constant

5.4What can be best concluded from 5.3?

a) Kinetic energy depends both on mass and velocity
b) <u>Potential energy depends on height</u>
c) Potential energy depends both on mass and height
d) Kinetic energy depends on height.

3.5.4 Item modifications

After the items were written by the researchers, they were given to six selected physics teachers who have been teaching grade eight physics subject to revise and incorporate new items. The teachers followed Haladyna et al.'s (2002) multiple-choice item-writing guidelines to validate the items and develop new items. The teachers provided comments on the progress levels and develop new items that were meant to represent the levels. Discussions were held with the teachers about the students' levels of reasoning ability and the suitability of the items for the students. Based on the teachers' comments and discussions the items were revised, and new items were incorporated (see Section 4.2). Based on the discussions with the teachers, twenty-two items were revised and five items were rejected and five new items were incorporated. The items were then being re-adjusted and contextualized to the students' reasoning levels and 44 items were made available for the experts.

3.5.5 Content validation

Beyond the participation of school teachers, two Ph.D. candidates with experience of teaching physics, specializing in physics education, and who took courses on assessment and scientific reasoning involved ensuring content validity. These experts were asked to check and evaluate the alignment between the items and the reasoning levels, categorize the items into three types of styles of reasoning, and check the quality of the items if the items represent the intended construct. The experts provided feedback related to the setting of the items, their alignment with the measure at each level, language, and conceptual problems along with the possible mechanisms on how to improve. The experts' comments and suggestions for improvement were duly considered when modifying the items during the second round revision. Out of 44 SRPT items reviewed, nine items were identified and revised to increase the content validity of SRPT (six as having vague wording while three as having scientific reasoning content). Through this process, 44 final items were made ready for pilot testing.

3.5.6 Construct validation and reliability calculation

One of the core tasks of this research is constructing and validating the Scientific Reasoning Assessment Tool. Hence, several activities were performed to ensure construct validity. These include using a standardized pool of items and incorporating experts' views discussed above, conducting pilot testing, and relating analysis results with theoretical bases.

Construct validity is said to be ensured when items measure a theoretical construct according to the predicted manner (see Section 2.10.2.6). Rasch analysis is used to establish construct validity by examining whether a given data represents the predicted theoretical construct (Bond & Fox, 2007; Cavanagh & Sparrow, 2010b; Martin & Jamieson-Proctor, 2019). Rasch analysis provides indicators that can help to determine the construct validity of a construct intended to measure. To ensure construct validity, this study employed Rasch statistics (mean square (MNSQ), Z-Standardized Scores (ZSTD), unidimensionality, item/person separation, and reliability indices), and representations of Rasch modelling (Item person map, bubble chart, and ICC curves). The first approach of Rasch modelling to ensure the construct validity is determining the quality of the items. When the all items fall in the expected range (fit to the model), the test is considered to be unidimensional or measuring a single construct. Fit statistics provided by MNSQ and ZSTD values help to determine whether each item contributes to measuring the intended construct (Bond & Fox, 2015). The MNSQ and ZSTD are reported by infit and outfit values. The data that fall in the expected interval for the indicators by infit and outfit values are likely to represent the underlined construct.

Unidimensionality is an attempt to measure a physical attribute by focusing on a single attribute or dimension at a time. MNSQ and ZSTD are for individual items whereas unidimensionality is about determining whether a set of test items as a whole contribute to measuring a single construct. Unidimensionality of the SRPT was detected with the help of principal component analysis (PCA) (see Chapter 5). Furthermore, the Rasch model relates item difficulty and person ability in a way that can represent a given construct (Bond & Fox, 2015). It provides a way to determine how items are hierarchically distributed in a continuum and persons are classified according to their achievement to represent an underlying construct. This is demonstrated by item/person reliability and separation indices. According to Boone et al. (2013), item separation with item reliability provides one to determine whether item hierarchy is in a predicted continuum in relation to person sample to ensure construct validity; person separation with person reliability provides one to determine whether an instrument distinguishes high and low achiever.

The second approach of Rasch analysis to ensure construct validity is determining the quality of the items whether they represent the underlying construct in a predicted order which involves a qualitative interpretation of a construct (Senocak, 2009). This can be achieved by the pattern of item distributions in terms of item difficulty and person ability using the item person map and bubble chart. Item person map is an important Rasch representation that depicts an illustration of students' ability and item difficulty of an underlying construct (Bond & Fox, 2015). It can easily visualize the way how items and persons are distributed on a single scale. It tells the gap between items, overlapping of the items, and the order of items in relation to a prediction made by the tool developer or researcher. The bubble chart is also the Rasch representation that can help one to determine items that fall in the expected range. It provides a representation of item difficulty in the ZTSD range of -2 to +2 (Bond & Fox, 2007).

This study employed the aforementioned indicators of Rasch analysis to determine the extent to which the SRPT measure the construct scientific reasoning. The tool, in turn, was used to determine students' scientific reasoning ability in terms of the established reasoning progress levels (see Chapter 5).

3.5.7 An example of how the items were developed and revised by using the assessment development procedure

The example below depicts how the items were developed using the assessment procedure (see Section 3.5) and person-item map (See Figure 4.1). This section provides how question 8 which consists of items 8.1, 8.2, 8.3, and 8.4 was developed, revised, administered for the students, and content and construct validity were ensured.

Step 1: Conceptual Framework

Conceptual framework was developed by considering styles of scientific reasoning, Ford and Wargo (2012), McNeill and Krajcik (2011) and Styles of Scientific Reasoning (see Section 2.4).

Step 2: Development of Construct map

Firstly the item map was generated as indicated in Table 4.10 and stduents' views were explored as indicated in Table 4.12. With the help of literature review, students' view towards the four levels of scientific reasoning, the person map was developed as depicted in Figure 4.1.

Step 3: Item Formulation

After the core objectives (specific outcomes) were identified for each levels of scientific reasoning, the items that can represent each level were developed. For example, Item 8.1 was meant to measure level 1, item 8.2 was developed to represent level 2, item 8.3 was developed to represent level 3, and item 8.4 was constructed to represent level 4.

Level 1: Identify the units of measurement:

Question 8: Two students perform an experiment. To accomplish this, they collect different working batteries, wires, and ammeter. After connecting properly they measured the values of current (I) in the circuit by varying number of batteries for a single value of resistance (R). The students' measurement is tabulated as shown below:

Voltage (v)	Resistance (Ω)	Current measured by first student (A)	Current measured by second student (A)
1.5	2	0.73	0.76
3	2	1.40	1.45
4.5	2	2.20	2.20
6	2	2.95	3.05

8.1 Which of the following is used to measure voltage?

a) Volt b) ampere c) ohm d) ammeter

Level 2: **Explain variable relationship**:

8.2 Which statements explain what students want to know from this experiment?

a) **They want to know the relationship between voltage and current**
b) They want to know how resistance affects current flow
c) They want to know the relationship between resistance and current
d) They want to know how current affect voltage

Level 3: **Generate an evidence** for the explanation in item 8.2:

8.3 Why both students did not exactly get the same measurement in the table?

a) There is some problem in measuring devices

b) <u>Measurement can't always give exactly the same value</u>

c) One of the students is not skilful in measurement

d) Both students are wrong

Level 4: **Draw a conclusion** based on the items 8.2 and 8.3:

8.4 What would you conclude if students use 7.5 volt battery in the above experiment?

a) The value of the current measured by the students will vary in the range between 2.95 and 3.05A.

b) The value of current measured by the students will be greater than 2.95 A

c) The value of current measured by the students will be greater than 3.05A

d) The measured value of current will be greater than 2.95A and 3.05A

Step 4: Content validation

Comments and suggestions were received from the experts and were considered when modifying the items. (See Sections 4.2 and 4.3).

Step 5: Field Test

After review of experts items were administered to the selected sample.

Step 6: Construct Validity

To analyze the data one dimensional Rasch's dichotomous model analysis was used. Using the Winstep3.68.0, the quality of the items was analyzed with the help of indicators such as separation index, item fit, correlation, unidimensionality, and item person map.

After the first version of SRPT administered to the selected sample, the data were analyzed using Rasch analysis, and misfit items (items that don't meet standardized value) were identified. To modify the items careful inspection was made to determine the nature of the misbehavior. There were items too complex to understand for the students, there were items suspected to measure other constructs than predicted in the theory, there were items that don't represent the indicated levels, and there were also items that lack clarity in writing. Accordingly, the items were re-written and the final version was provided in the Appendix A.

Step 7: Judgmental stage

At this stage items were accepted, modified or rejected and replaced by other items in light of aforementioned steps (see Sections 4.2 and 4.3).

3.6 The Pilot study

According to Bless, Higson-Smith, and Kagee (2006), a pilot study is a small study conducted prior to a larger piece of research to determine whether the methodology, sampling, instruments, and analysis are adequate and appropriate. For this study, piloting was conducted to evaluate the quality of items, check the consistency between items and the construct map, check the relevance of the framework, make a necessary revision, and determine the duration of time required for the test. The pilot test was administered to grade eight students of Kambata Zone, SNNPR.

To analyze the data, one dimensional Rasch's dichotomous model analysis was used. Using Winsteps 3.68.0 software (Linacre, 2020), the quality of items was analyzed with the help of indicators: separation index, item fit, correlation, unidimensionality, and item person map.

3.7 Data collection and Data Analysis

For this study, the data were collected at different phases. In the first phase data were collected qualitatively to explore students reasoning pattern, to develop the construct map and the tool, and followed by quantitative data which served as validation of the construct map and the tool.

3.7.1 Qualitative data collection

To answer the research questions 1 and 2 data were collected qualitatively. Twelve students were interviewed by the researcher and audio recorded. It was conducted in Amharic. The interviews were transcribed verbatim from the audio record hearing the audio repeatedly. The interview questions were developed in a way to explore students' way of reasoning for the four levels of reasoning established from the literature: factual knowledge, explanation of factual knowledge and evidence-based reasoning, and drawing conclusion. FGD was also conducted with six primary school teachers to further corroborate the result of students' interviews and to validate the tool developed by the researcher. Furthermore, to ensure the content and construct validity of the SRPT panel of experts were participated by reviewing the items of SRPT.

3.7.2 Qualitative data analysis

To develop the categories of information, the data were sorted and grouped iteratively. To generate meaning and categorize into groups the words, phrases, expressions, and examples students provided were recorded, sorted, and re-sorted. Through the iterative process of the analysis, three themes emerged from the data for the four reasoning progress. The emerged themes were scientific views about physics concepts, scientific views with day to day practices (mixed views), and naive views about physics knowledge..

The analysis of the FGD with teachers also followed the same procedure. However, in this case, the FGD aimed to validate the results of the students' interviews. It was found that a similar result was obtained as students' interview with some additional points to strengthen the interview results.

3.7.3 Quantitative data collection

To collect quantitative data, the survey method was used. Following item development and modifications, the test was administered to a large sample. The quantitative data were collected in two rounds. In the first round (pilot-testing), the test was administered for 203 students at six selected schools of Kambata Zone. The test was analyzed and some items were modified, some of them remained as they are, and some are rejected, some new items were incorporated into the tool. The final test was administered for 242 students at the schools of Hadiyya Zone.

3.7.4 Quantitative data analysis

In this study, one dimensional Rasch's dichotomous model was conducted to evaluate the psychometric properties of the SRPT using Winstep 3.68.0. The quality of items was examined by the indicators of Rasch Modeling such as Item fit, Bubble chart, item person map, unidimensionality, item and person reliability, category functions, and item characteristic curves (ICC).

Item fit was used to examine how well the items fit the model and how the items behave in terms of standard values represented by mean square residuals (MNSQ) and standardized square residuals (ZSTD) values to represent the construct scientific reasoning. Unidimensionality was used to identify unexplained variance and determine whether the items measure unexpected construct than in the question, and how well the items work together to define the construct i.e. scientific reasoning. With the help of (principal component analysis) PCA, the variance was examined to identify explained and unexplained variances. Item person map was used to examine how the items and person were distributed concerning the item difficulties and person abilities pictorially. The Bubble chart was also used to define the direction of the underlying variable. ICC was used to determine the precision of the test in terms of the expected model (ideal curve).

3.8 Validity and Reliability

There are no rigid rules to ensure validity and reliability in a qualitative instrument (Creswell & Poth, 2018). However, in mixed methods, there is a need to have evidence of validity and reliability in a flexible way. Various procedures have been used to ensure the validity and reliability of the instruments developed and result obtained in this study. The instruments used for the interview and FGD were validated by two Ph.D. candidates for relevance and clarity. Based on the comments the researcher made modifications such as incorporating additional question, probing questions, language improvement, and structure of sentences (see Appendix D, E, and F).

The validity of the qualitative data was ensured with the help of four components: credibility, transferability, dependability, and conformability. To ensure credibility, the students' interviews were validated with the help of focus group discussion (FDG) with primary school teachers, and to ensure the FGD results, discussions were held with the teachers about the results. The SRPT was also reviewed by the experts. To ensure transferability, students' were interviewed until the needed data was saturated and purposive sampling was used to select teachers who have three years of participation and experiences in implementing activities related to scientific reasoning. To ensure the dependability of the result, the interview and FGD recordings listened to again and again, and two Ph.D. candidates were also participated in coding the interview transcripts. To ensure the conformability of the results discussions were held with the primary school teachers and two Ph.D. candidates in physics education at Addis Ababa University.

To ensure the face and content validity of SRPT, different experts were involved (see Section 3.5). The experts provided feedbacks related to the setting of the items, their alignment with the measure at each level, language, and conceptual problems along with the possible mechanisms on how to improve. Furthermore, to ensure validity and reliability of the tool pilot study was conducted (see Section 3.8) and the final tool was analysed using Rasch's dichotomous model that can provide different indicators.

3.9 Ethical issues

In this research primary school students, primary school teachers, college educators, and Ph.D. candidates were involved as the source of data. The researcher is required to state his position concerning his role in data collection, analysis, and interpretation. There is a need to ensure that the participants were involved voluntarily, and all the necessary ethical standards have been maintained.

The procedures the researcher followed in this study were not harmful to participants' moral and ethical values. The participants' rights were not violated. In all aspects, the study followed legal procedures. Respect was given for participants; that means protecting autonomy, ensuring well informed and voluntary participation were secured. The study did not put any of the participants at risk psychologically, physically, legally, socially, or economically (Creswell & Poth, 2018). The purpose of the study was described for the participants before they took part in the study. The students' and teachers' names were not recorded, the report was provided by using pseudonyms, and none of the information collected from the participants was used to identify the participants as individuals. The purpose of the study was informed to the school administrators and teachers, and a letter of consent was given to the teachers and administrators (see Appendix H and Appendix I) before administering the test. The administrators and teachers have expressed their willingness. Finally, the data were analyzed and interpreted to provide an accurate account of the information through the validation process (Creswell & Guetterman, 2019).

This chapter discussed how the pragmatist paradigm provides educational researchers to mix ontological and epistemological assumptions as well as the methodological processes of the research to develop, validate assessment tools. It was also presented that how a sequential exploratory design firstly exploring students' patterns of reasoning using phenomenography (qualitatively) to develop SRPT which is followed by investigating students' levels of reasoning using descriptive survey research (quantitatively), to validate and investigate students reasoning pattern, was used in the current study. The chapter also presented the source of data, sampling, data collection, and analysis procedures. Finally, it was also discussed the assessment development procedure, the validity of the analysis, and ethical issues. The next chapter will present the qualitative result of the current study.

Chapter 4 Qualitative analysis and results

This chapter presents the analysis and results of qualitative data collected during the development and validation of the scientific reasoning instruments. The qualitative data were analyzed to answer the first and the second research questions. The first research question: “Which scientific reasoning skills are addressed by the construct map developed for the SRPT test items?” was tried to answer using a review of literature, students’ interviews, and FGD. The second research question: “How does the content and construct validity of the SRPT items qualitatively established?” was answered using discussions with the primary school physics teachers and a panel of experts.

4.1 Qualitative analysis

In this section, the qualitative analysis is described. Qualitative data were collected from the panel of experts, elementary school physics teachers, and grade eight students. The panel of experts gave their critical feedback about the items of SRPT. In this section, the result of students’ interviews and FGD with the teachers will also be discussed.

4.1.1 Analysis and findings of students’ interview

Analysis of interview transcripts provided insights about the nature of grade eight students’ scientific reasoning abilities. In addition to carefully listening to the audio of the interview, the researcher thoroughly and repeatedly read the interview transcripts. Then, repeated terms, examples, expressions, and phrases that are related to the nature of scientific reasoning across various levels were coded from the transcribed text for each level of students' reasoning. In the first cycle of coding essential utterances that depict the nature and characteristics of students' reasoning abilities were identified. Then in the subsequent coding stages, these views of students’ reasoning were condensed, and eventually, resulted in three major themes relevant to students' scientific reasoning experiences. Two Ph.D. candidates, from the Department of Science and Mathematics Education at Addis Ababa University, participated in coding the interview transcripts. Differences in coding between these coders were resolved using negotiation. For the sake of validity, this study quotes verbatim taken from the students' utterances of the interview transcripts. Moreover, interpretations of utterances of grade eight students' scientific reasoning experiences will be provided.

As discussed above, analysis of grade eight students' interview transcripts resulted in three main themes regarding their level of scientific reasoning. These three main themes are *naive, mixed, and scientific views*. The interview transcripts were coded using In Vivo coding during the first cycle of coding (Saldana, 2013). Then, the second cycle of coding of students' interview transcripts deployed axial coding to develop and to gather the In Vivo coding results into themes. When student utterance was considered as experiential and intuitive, the utterance is coded under the theme *naïve views*. For example, utterances such as *facts are truth, facts are certain, facts are absolute, etc.* were categorized as *naive views.* When a student utterance was aligned or matched with the accepted scientific views, the utterance is labeled as *scientific views.* From students' responses, one can see that there are scientific views about physics facts such as *facts in the textbook are important to construct scientific knowledge*. Such a view about scientific knowledge is termed a *scientific view*. When student utterance mixed intuitive and accepted scientific views, the utterance is interpreted as *mixed views.* Students' responses such as *physics facts correctly represent a natural phenomenon, facts are to be remembered...* shown mixed students' views about physics knowledge. *Mixed views* amalgamate *scientific views* with *naïve views.* Often, students, who possess *mixed views*, combined naïve theories with scientific theories.

To sum up, *naive view*s stand for students' reasoning based on the non-scientific, intuitive, irrelevant, and incorrect responses; *mixed views* refer to students' reasoning based on the students' mixed up responses of intuitive practices with scientific knowledge; *scientific views* refer to students' reasoning that agrees with scientific knowledge and theories.

Six guiding questions (see Appendix E) were developed to identify the abovementioned themes and explore the nature of students' scientific reasoning in physics. Based on the responses of students' three themes, naïve, mixed, and scientific view, emerged in each level of reasoning, namely *factual knowledge*, *explanation of factual knowledge*, *evidence-based reasoning*, and *drawing conclusion*. Hence, these three themes of students' reasoning were explored concerning students' reasoning for each of the four-level of reasoning. In the next section, a detailed description of students' interview with the help of excerpts is presented.

Under each level of scientific reasoning, students' patterns and progress of reasoning are discussed.

4.1.1.1 Factual knowledge

The question that was asked for this level was:

> How do you view physics facts and concepts? (*Probing question: Do you think the definitions, constants, formulas…in the textbook are truths to be grasped or something needs to be checked, proved, and supported with evidences? Give one example from physics concept*)

The analysis of this interview question indicated that respondents possessed different views about facts and concepts in physics. Four of the interview participants (33%) of this study indicated that they had *a naïve view (see Table 4.2)* about the nature of facts and concepts in physics. Students, with *naïve view,* assumed that scientific facts, physical theories, and natural laws as identical terms about the laws of nature. In addition to this, students, with *naïve view,* believed that established physical theories and laws are immutable.

It seems that these students took scientific knowledge as absolute and guided by an infallible principle of nature. This conception of students created difficulties in their attempt to generate scientific knowledge. Four of the respondents (33%) showed *mixed views*. Students, with *mixed views*, took scientific knowledge claims written on a textbook or told by their teachers as a representation of scientific '*truth*'. Perhaps, students, with *mixed views*, may accept such knowledge claims based on authority. Students are not skeptical of their teachers and textbooks. Here, students in *the mixed views* category focused on understanding the product of science, i.e., factual knowledge not on understanding the processes of science that produced the factual knowledge. In *the mixed views* category, students focused on recalling facts and rote memory rather than providing explanations about physical phenomena and relations between concepts of physics. *Three* students' (33%) had shown *scientific views* in their responses. Students, who possess *scientific views*, thought that the collection of dependent facts are the basis to build knowledge. Facts or a combination of facts is required to develop a particular understanding. Facts are the beginning of understanding physics in a new way. Hence, acquiring facts is not the goal of learning physics. Facts are conditional expressions of physical quantities and physical phenomenon. Students, with *scientific views*, believe that scientific facts and physics theories need to be considered as interconnected. Table 4.1 presents three excerpts that are taken from students interviews.

Table 4.1 Excerpts taken from students interviews about factual knowledge.

Excerpts taken from students interviews	Codes
Physics deals with physical quantities which are defined and represented by simple facts. I believe the facts, concepts, and formulae in the text books are important to understand natural phenomenon but we need to check for its relevance. (Student 1)	**Scientific views**
Physics facts are observations and concepts discovered by scientists. They are reliable knowledge to be learnt and understood. That is why they are presented in the text books and teachers are engaged in teaching for students. (Student 12)	**Mixed Views**
I believe all the concepts written in the textbooks are truth value. All the physics concepts presented in the textbooks are tested and experimented before decided to present for the students' learning. (Student 4)	**Naïve Views**

The response of student 4 is an exemplar of naïve views. Students, with naïve views, believe that scientific facts are truth. According to such students' beliefs, facts are concepts, laws, and theories discovered by physicists to be recalled and remembered. Once knowledge is discovered by someone it works forever, according to some students. That is, a teacher needs to catch that revelation for his students and the students' responsibility is attempting to catch that reality. Such views seem to be an obstacle for students' high-level reasoning and knowledge constructions.

Respondent 12 is under the mixed view category. It was revealed from such students' responses, there are students who think that written facts are not truth; they are subject to change but need to be taken as they appear. Even if they understood that physics facts are subject to change and socially constructed, they struggled to use facts as a foundation for the construction of other levels of understanding and reasoning. They think that facts are the end product of physics knowing. Such views might hinder students from developing higher reasoning abilities. For such students learning physics means understanding what teachers tell them so or what is written in the textbooks. According to such a view, it is enough to follow the teacher's note and re-read what is written in the textbook. Such students do not bother about reasons but they make meanings based on notions and feelings. For such students, there is no difference between speed and velocity, distance and displacement, temperature and heat, energy, and work.

Respondent 1 is under the scientific view category. Students' with scientific views believed that facts are mere expressions to represent the physical occurring phenomenon. This seems a proper view of scientific facts. This view can be progressed when students develop the abilities to engage in discussions by using scientific facts to explain their claim. These students think in advance considering facts as a cornerstone for constructing and discovering up-to-date knowledge, and higher reasoning level.

Table 4.2 Students' view about factual knowledge

Factual knowledge	Students achieved the level (number)	Students achieved the level (%)
Naive	4	33
Mixed	4	33
Scientific	4	33

4.1.1.2 Explanation of factual knowledge

The question that was asked for this level was:

> How do you view scientific explanations? *(Probing questions: how scientific explanations are constructed? Is there any relationship between scientific explanations and day to day experiences? Give an example from physics concept.)*

The interview of grade eight students' responses indicated that five students (41%) had a naive view (see Table 4.5) about this level of reasoning. Students who had naïve views believed that the law of physics is the law of nature. They think that physics law, theories, and principles are infallible. For such students, providing explanations is discovering naturally aligned working principles for various things in common. The excerpts of Student 4, which was given below, reflected such perspectives. Based on this naïve view, the working principles of various discoveries are something that has already been existed but hidden because of lack of knowledge. This view about physics knowledge can't be expected to produce scientific reasoning.

Table 4.3 Excerpts taken from students' interviews about explanation of factual knowledge.

Excerpts taken from students interviews	**Codes**
Physics, for me, is a subject which tells about the phenomenon which occurs in our surroundings. It gives description for things we observe and practice in daily basis. For example, when we take the concept velocity I can understand its meaning with the distance covered from starting to end point in a specific direction by measuring time it takes to cover the path length. (Student 1)	**Scientific views**
After I learn Newton's second law of motion I understand why a moving body comes to rest. I now understand that if there is no force there is no motion. Object move from one place to another because they posses force within them. (Student 11)	**Mixed Views**
One needs to give meanings for a physical phenomenon based on understanding of how the constituent elements of that phenomenon work. Unless one discovers and understands how things are aligned naturally, he can't give correct relationship between them and the explanation doesn't work anymore. (Student 4)	**Naïve Views**

Students' responses revealed that four (33%) students believed that learning physics concepts have nothing to do with their day to day activities. However, after the physics concept is introduced they tend to create meanings from what they see, touch, and do regularly. As far as the concepts are not provided in terms of concrete and tangible things, they remain meaningless expressions in their minds. Even if students at this level are willing to construct their mental theory about nature, they struggled to construct scientifically. They struggle to relate physics concepts with day to day activities scientifically or they try to relate physics concepts with day to day experiences wrongly. Excerpts of Students 11, which showed a mixed view, is given above. Student 11 seemed to mix up science concepts with daily experiences based on a wrong understanding of physics knowledge. Such students' view and way of explanation is not expected to produce scientific reasoning because they incorrectly interpreted the law of physics in the first place.

Scientific views for explanation of factual knowledge were observed in 3 (25%) students' reasoning (see Table 4.4). After students having an understanding of facts to some extent, they tend to relate different physics concepts to explain a physical phenomenon. Using facts as a basic element to construct knowledge, they start to give meanings to physical quantities. This was explained by Student 1 as shown above. This indicates that the student is trying to understand the concept of velocity by relating the concepts of distance, time, and direction. The concept of velocity for him is abstract until he creates a relationship between the physical quantities distance, speed, and time with direction. This way of explaining concepts leads students to help in promoting higher-level understanding and reasoning.

Table 4.4 Students views about explantion of factual knowledge

Explanation of factual knowledge	Students achieved the level (number)	Students achieved the level (%)
Naive	5	45
Mixed	4	33
Scientific	3	25

4.1.1.3 Evidence-based reasoning

The question that was asked for this level was:

Does the construction/development of scientific knowledge require evidence? If so, what are the sources of evidence? (Give an example from physics concept)

Most views (8 students or 58%) about evidence (see Table 4.5) tend to maintain the prior beliefs or to arrive on the knowledge of the law of nature assuming the law of nature is to be manifested. (See excerpts of Student 2 provided in Table 4.6). For such students, physics knowledge is infallible truth already proved by scientists. They thought that the law of physics is the same as the law of nature. Based on this belief the only thing required from students is trying to catch what has already been discovered. This also indicates that the importance of evidence is only to establish and apply the knowledge which was discovered by scientists in practical aspects and to arrive at unchangeable knowledge. A sample of stduent response was provided which strengthen such view about evidence. (See excerpts of Student 6 provided in Table 4.6Table 4.6).

Table 4.5 Students views about evidence based reasoning

Evidence-based reasoning	Students achieved the level (number)	Students achieved the level (%)
Naive	8	58
Mixed	2	16
Scientific	2	16

Table 4.6 Excerpts taken from students interviews of evidence based reasoning.

Excerpts taken from students interviews	Codes
Evidence is required because we can't be sure whether a claim is correct unless we have proper evidence. Therefore, one needs to check whether a claim works or not before accepting or rejecting. Even the physics concepts, laws and theories are man-made, theoretical and subject to change and modifications. (Student 1)	**Scientific views**
One needs evidence to understand scientific concepts in a better way. The sources of evidence are teachers, experienced persons and natural phenomenon. Evidences can be generated from individuals who have better knowledge and understanding in physics. The evidence for the presence of gravity is falling objects. The evidence for the presence of magnetic force is repulsion and attraction of magnetic materials. (Student 2)	**Mixed Views**
Evidence is needed to be generated from physics laws and also from practical demonstrations. We need evidences to understand the law of nature. Scientists have already discovered the laws of nature but as a student I need evidence to arrive on the knowledge discovered by scientists. The importance of generating evidence is not to accept or reject the findings of physics but it is only to improve my understanding about physics laws and theories. (Student 2)	**Naïve Views**
It is not required to prove or disprove what has already been discovered by scientists. The concepts are already tested and proven to be true by scientists. The findings or theories are to be accepted and learnt to use them in our daily life. Evidence is not required to reject or accept scientific findings. Evidences are only for the sake of understanding what has already been proved by scientists. (Student 6)	

Students (two students or 16%) have also shown the tendency to generate evidence for claims by relating day to day experiences with physics laws (see Table 4.5). They believed that evidence is required to construct knowledge. Any observations, experiences, demonstrations, and examples are considered valid evidence to support a claim. They are also convinced that the most accurate evidence can be generated by experienced persons, good teachers, and physical phenomenon. (See excerpts of Student 2 for particular responses). This view indicates that as far as evidence works for a particular situation it could be considered as valid. Students' seemed to be heavily dependent on external bodies and practical demonstrations to generate evidence. Students believed the significance of evidence but there is no clear view about scientific evidence. This way of viewing evidence implies that there is a challenge in identifying valid and invalid evidence.

There were a few students (two students or 16%) who thought that evidence is required to construct new knowledge and also to accept or reject a claim. They understood that scientific knowledge is dynamic, ever-expanding, and subjected to both modifications and changes. They knew that science laws and theories are man-made and have limitations. (See excerpts of Student 1 for such explanations). Such students understood that scientific evidence is required to accept, reject, modify, and develop new knowledge. They also believed that there are valid and invalid evidence based on the relevance to support the claim.

4.1.1.4 Drawing conclusion

The question that was asked for this level was:

Can we arrive on a certain conclusion based on the evidence? (Probing question: How can one draw a scientific conclusion?)

Most students (9 students or 75%) believed that the law of physics is the same as the law of nature (see Table 4.7). According to such beliefs, there are no personal conclusions to be claimed scientific. Conclusions for such students are to arrive on an ever-working principle of nature for a given phenomenon. They concluded that, if someone arrives at a conclusion about a phenomenon based on valid evidence, there would be no change for such a finding. (See excerpts of Student 3 for such particular responses).

According to such students' beliefs, scientific conclusions are reserved for scientists and the ones who uncover the law of nature. The others' responsibility is only attempting to understand and apply in life situations. One can construe from this belief that scientific conclusion is the final say about the working principle of nature.

Table 4.7 Students views about drawing conclusion

Drawing conclusion	Students achieved the level (number)	Students achieved the level (%)
Naive	9	75
Mixed	2	16
Scientific	1	8

Some students (two students or 16%) believed that conclusions are based on the experiences they observe from the environment (see Table 4.7). They attempted to draw a conclusion by mixing-up scientific theory with day-to-day activities. When students are asked to relate about a physical phenomenon based on the physics laws and theories, they give wrong generalizations. Such students claim that they draw a conclusion based on scientific knowledge but there exist contradictions with the scientific findings. (See excerpts of Student 5 for such particular reflections). From the response of Student 5, one can construe that the student struggled to draw a scientific conclusion. The reason for this could be the lack of correct scientific understanding of the concept of relativity. It seems that he used the law of physics to arrive at a wrong conclusion. It suggests that a lack of detailed knowledge about physics laws and theories might leads students to relate scientific knowledge with day-to-day practices to arrive at a wrong assumption, prediction, and conclusion.

Table 4.8 Excerpts taken from students interviews of drawing conclusion.

Excerpts taken from students interviews	Codes
We can generalize based on the available evidence, we can reject or accept based on the experimental tests but we can't arrive on a final and ever working conclusion. The understanding of physics concepts, theories, and laws are ever expanding and subject to modifications. (Student 1)	**Scientific views**
One can arrive at a conclusion based on the physics laws and theories. One can draw conclusion about a physical phenomenon by relating physics knowledge. For example based on Newton's first law of motion plants are at rest and will remain at rest for ever. They are always at rest unless someone moves them from where they are. According to this law there are objects which never move for ever. (Student 5)	**Mixed Views**
Scientists and philosophers draw conclusions after several trials and experiments. The laws and theories are accepted as truth only when they are proved and checked the alignment with the law of nature. But one can generate available evidences to improve his knowledge before arriving on the conclusions scientists made. (Student 3)	**Naïve Views**

There was a student (8%) who claims that one can arrive at currently working conclusions based on the evidence, discussions, and arguments. For such students, scientific knowledge can be modified, rejected, and accepted depending on the quality of evidence. Such students believed that no one is exempted from making mistakes; there is no absolute reality, and every finding should pass a test of time before application and taken as scientific knowledge. (See excerpts of Student 1 for such particular reflections). Even if they were a few in number some students believed that based on the valid and reliable evidence one can arrive at a tentative conclusion and they also understood that there is no ever working scientific knowledge.

Analysis of students' interview transcripts resulted in further elaborations about grade eight students' nature and character of scientific reasoning. In the above section, a description of students' reasoning across the four levels of reasoning was provided. In what follows a summary of the analysis of interview questions is given (see Table 4.9).

Table 4.9 Summary of students' interview

Reasoning pattern	Sample transcripts	Descriptions
Naive	...9.8m/s^2 is truth value... (fact) (student 3) F = ma is law of nature discovered by scientists (explanation) (student 5) ... ...physics laws and theories are true evidence (evidence) (student 9) ......findings are revealed truth (conclusion) (student 11)	✓ Physics facts are truth, universal, and certain ✓ There is universal explanation ✓ Physics ideas, concepts, laws and theories are the source of absolute evidence ✓ There is true/absolute conclusion (conclusion)
Mixed	...facts are observations and concepts discovered by scientists. They are reliable knowledge...(fact) (student 12) ... velocity represents how fast a body moves (explanation) (student 6) The sources of evidences are teachers, experienced persons and natural phenomenon... (evidences) (student 7) ....based on Newton's first law of motion trees out there are at rest and will remain at rest for ever. (conclusion) (student 5)	✓ Mixed understanding about physics concepts ✓ Unable to differentiate between speed and velocity ✓ unable to identify valid and invalid evidences ✓ Unable to draw scientific conclusion
Scientific	...100 km represents a distance... (fact) (student 8) Velocity......is the distance covered from starting to end point in a specific direction.... (explanation) (student 12) The evidence for the presence of gravity is falling objects (evidence) (student 1) ....F = ma can be improved (conclusion) (student 1)	✓ Facts are foundations to construct new knowledge. ✓ Provide an explanation of variable relationship... ✓ Evidences can be generated from physics concepts, laws and theories ✓ Understand tentative nature of conclusion

Findings from the overall students' interview revealed that the level of students' scientific reasoning was dominated by factual reasoning and their views were also dominated by naïve theories. Most students mentioned their conception of facts, laws, and theories in physics. Their discourse was dominated by ideas such as *Physics facts are true, universal, and certain; explanations are universal; physics ideas, concepts, laws, and theories are the source of absolute evidence;* and *there is true/absolute conclusion*. Such responses revealed that the majority of students assumed scientific knowledge as static, fixed, universal, certain, and unchangeable. These students perceived physics as mere collections of facts, concepts, laws, and theories to be grasped instead of thinking and realizing physics as a culture of practicing and exercising activities to construct tentative knowledge. Beliefs such as scientific knowledge is unchangeable truth might push students' science learning towards recalling of physical laws and rote memorization of facts. Student interviews provide insight into the nature and characteristics of students' scientific reasoning. To validate the students' pattern of reasoning and their views in terms of the four scientific reasoning levels the subsequent section dealt with teachers' FGD.

4.1.2 Analysis and findings of Focus Group Discussion

The researcher facilitated the focus group discussion (FGD) with the six selected primary school teachers using guiding questions developed for the students' interview to explore students' reasoning pattern under each theme that emerged from the students' interview. The discussion was held in the physics laboratory room at Hossaena College of Teacher Education. The place was selected for its quietness which is appropriate for such a discussion. All the participant teachers engaged actively in the discussion. The purpose of the FGD was to discover the alignment, consolidate, and validate the findings obtained by students' interviews. FGD, here, served as a tool to explore more information about the themes which have already been emerged by students' interview.

The primary school teachers reflected their own beliefs, understanding, and experiences about students' reasoning during the discussion. The teachers were selected for this task because they were participated in the TPSS research project**.** TPSS was a three-year research project which aimed for improving the quality of teaching in STEM subjects in Ethiopia. During this period the teachers took training on dialogical teaching which engages students in discourse with a teacher and other students, scientific argumentation and reasoning, developing scaffolding activities for evidence-based discourse, and implemented the activities in the classroom for a year. The teachers were also provided with a material that contains various activities on scientific reasoning and argumentation, and that can guideline how to implement the activities for the students.

The data were recorded using an audiotape recorder. During the discussion, the primary school teachers identified common features of students’ scientific reasoning such as naïve reasoning, alternative conceptions, disorganized ways of thinking about physics laws, poor sources of evidence, and unscientific ways of drawing conclusions.

The analysis was done as described in Section 4.1.1 above. Firstly the researcher listened to the recordings of FGD again and again, and read and re-read the transcripts to select segments of texts. Secondly, the data were interpreted in terms of the selected segments of texts, words, expressions, and examples by listening to the recordings, and transcripts, again and again, to give meanings for the teachers' responses and develop themes. Finally, the interpretations were coordinated, the three themes emerged during students’ interview were also appeared.

The Table 4.11 demonstrates how valid was the themes emerged from the students’ interviews and how the similarity appeared with the students’ results. The following section presents the results obtained from the focus group discussion with teachers.

4.1.2.1 Factual knowledge

4.1.2.1.1 Scientific view

It was found from the teacher’s responses that students understand that facts are important to construct a higher form of knowledge and to develop reasoning. Students at this stage emphasize facts beyond memorization instead as a necessity to develop a higher level of knowing and thinking. For this reason, students try to differentiate one concept from the other, one unit measurement from the other, one fact from the other. One of the teachers put this as:

This is very important level for students. They give more emphasis for this level than other levels because they realize that without factual knowledge they can't explain, solve problems and give reasons for their claims. (DD)

This indicates that students at this level use factual knowledge as a transition to a higher level of knowledge and reasoning. This suggests that students have a scientific view about the level of factual knowledge which is essentially for all the higher level of understanding and reasoning with evidence.

Unlike the students' interview, here, the mixed view at factual level was not clearly emerged. This might be because students' mixed view is more related to explanation level for the explanation seeks one's understandings. At the explanation level, students' mixed view was clearly emerged in both students' interview and FGD than any other levels.

4.1.2.1.2 Naive view

It was found that low-level reasoning is characterized by students' perception of factual knowledge. Factual knowledge was considered as all about learning physics. In this level, the factual knowledge such as defining terms, remembering constant values given in the textbook ($g = 9.8\ m/s^2$, room temperature is 25^0c…), and remembering formulae are considered by students as reality. Students believed that learning physics aims to remember the facts. Memorizing the facts taught by teachers and the facts written in the text is considered as the end product of learning physics. The indication raised by the teachers for this belief is that students give more focus to take written notes on the blackboard by teachers than engaging themselves in various activities. Most of the time students don't bother about participating themselves in the discussions than copying what is written on the blackboard. As MK3 tried to explain this as:

When I ask questions students try to open exercise books or text book to answer the question thinking that the facts in the text book are truth or correct answers. Some other answer questions from their memory of what they read and write in their note *books. They also consider that the constants given in the text book is true value. (MK3)*

This tells that students in this level struggled to accept anything outside written facts, definitions, and formulas. Anything outside written facts, definitions, formulae are considered as errors. They like external 'truth' outside themselves. The factual level is characterized by students' responses without understanding and pondering; they respond as it appears in their mind.

4.1.2.2 Explanation of factual knowledge

4.1.2.2.1 Scientific view

From the teachers' responses, it was revealed that after students are introduced to physics concepts they face challenges to understand the relationship between physics concepts and prior knowledge. They then start to understand that physics is not abstract as they were expected before coming to the class. It was also construed from the discussion that students who have better knowledge and understanding at the factual level, explain concepts by providing related examples. Based on the knowledge they attain from the subject and prior conceptions, they answer questions by giving their meanings. One of the participants, MUK3, sated this as:

> *Students reasoning begin by comparing what they were told with what they learn in the classroom. Students start relating what they know before the concepts are introduced to what they learnt in the classroom. After students are exposed to physics concepts they start looking for pattern, create link between known and unknown to construct their own view*. (MUK3)

This suggests that one of the reasons that compel students to attempt providing explanations is the uneasiness created in their minds because of contradictions between what they knew before and after new knowledge is introduced. A few students ask questions boldly to resolve the problem of contradictions. They finally provide scientifically acceptable explanations.

4.1.2.2.2 Mixed view

The beginning of students' reasoning is trying to create a link between physics concepts with day to day activities before providing scientifically correct explanations and seek evidence. At this stage, they try to establish a temporary mind formulation (accumulation of meaningless information) about physics concept in terms of day to day experiences. This was explained by one participant, DD, as:

> *Students' eyes start to open when they are asked to provide their own understanding about their claims. This time they begin to see thing seriously. They try to understand similarities, contradictions, relationships between things... this is the beginning of establishment* of their own knowledge and understanding. (DD)

This level is characterized by students' attempts to give meanings for natural phenomena such as the motion of objects, the motion of charges in wires, motion of particles, transfer of heat. At this level, they reflect their view and position but most students fail to provide evidence that supports the claim. When they are asked the reason for their claim, either they prefer not to say anything or they provide unrelated evidence unconsciously. Only those who are successfully able to integrate temporary conceptualizations by creating a relationship between the prior view and newly emerged views start to seek evidence to consolidate or reject the mental theory. Even if students' at this level are willing to construct their mental theory about nature, they struggled to construct scientifically.

The level of explanation requires students' great struggle because it is the level that determines the way they see things in a new glass and seek to develop personal mental theory about the law of nature (Keys, 1999). It also determines students' progress of reasoning towards a high level. The level is characterized by accepting, modifying, or rejecting previously informed or known ways of reasoning. The success of the level is determined by the quality and amount of evidence they are confronted with. This level requires a special need for it determines the future of students' reasoning and learning physics. This requires providing contextualized activities and facilitation for the successful students' reconstruction of a new way of reasoning.

4.1.2.2.3 Naïve view

At this level, students try to explain things purely based on the knowledge previously informed. They try to explain based on the view they have for physics. The view for physics laws, theories, and knowledge is absolute and ever working. Based on that view they attempt to be perfect and always seek correct explanations. Such a view prohibits students to try and develop genuine knowledge. One of the participants reflected this as:

> *Some students are not bold enough to share their views, explanations and constructing ideas freely. They, most of the time, are not bold enough to share their understandings for fear of mistakes. They respond from the view of a fixed mind set.* (TK4)

It was construed from the discussion that students tend to make use of science concepts to establish prior belief without catching the very meaning of the concept. They were not ready to accept scientific conceptions against prior knowledge in the beginning. They don't allow accepting the scientific conception with open-mindedness to re-adjust prior conception instead they try to say what the concept doesn't say. This was the same finding as it was found before five decades earlier that most Ethiopian students persist to maintain unscientific beliefs (Woodward, 1969). Students face difficulty in accepting scientific findings when the observations do not much with the prior belief (Chinn & Malhotra, 2002). These authors also suggested that such problems can be overcome by devising proper scaffolding.

This is also in agreement with literature which insists that children construct mental theories to understand the world starting from early ages, but these theories need to be revised and evaluated when they grow from middle childhood to adolescence by engaging in an evidence-based dialogical discourse (Kuhn & Dean, 2005). Different research also indicates that students' reasoning can be affected by prior domain-specific knowledge and contextual knowledge (Kuhn, Schauble, & Garcia-Mila, 1992).

4.1.2.3 Evidence-based reasoning

4.1.2.3.1 Scientific view

It was also found from the discussion that students provide evidence from physics concepts, laws, and theories they learned before. But the challenge is the inability to create a meaningful link between evidence and claim. This is also illustrated by one participant, TK4, as:

> *Students provide evidence from physics laws and theories. But they face difficulty to clearly match the situation with exact laws and theories. Their evidences, most of the time, do not have clear alignment with the claim. (TK4)*

It was revealed from primary school teachers' responses that their students lack the ability to justify their reasons and to use data as evidence to defend their claim or explanation. Moreover, their students did not examine the relevance and sufficiency of their evidence. Only a few students are observed to be engaged in generating valid evidence.

4.1.2.3.2 Mixed view

It was found from the discussion that once students can explain ideas by their understanding, there is a trend of trying to generate evidence for their claim. At the explanation level, students try to develop conceptual theories in their minds about the relationship between physics concepts and day to day activities. Students develop an interest to verify the temporarily created mental assumptions using practical evidence. The evidence of students is characterized by trying to put things into practice. The students' theoretical mental construction about the relationship between science concept and day to day practices needs to be stabilized with practical evidence before drawing a conclusion. One of the participants, MUK3, reflected in this regards as:

> *Students prefer to generate evidences from practical aspect than physics laws and theories. Evidence for students is something which can be seen practically, to be demonstrated, and works in day to day activities. Without practical evidences, students' mental construction about physics concepts will remain illusion as they have already expected before coming to class. (MUK3*)

It was construed from the teachers' response that students provide practical evidence that fits with the claim, some other provide incomplete evidence and some other also provide incorrect and irrelevant evidence. Evidence provided by students is very similar to scientific but mixed with their prior beliefs and daily experiences which need to be treated and resolved by contextualizing the activities. They face difficulty to merge their evidence with the scientific one. It was stressed in the discussion that such a contextualization requires teachers' higher level of conceptual and pedagogical knowledge to help students assimilate the concept.

4.1.2.3.3 Naïve view

Teachers also revealed that students provide evidence that is not related to the issue raised during the discussion. They simply generate for the sake of providing evidence. This implies that there are students who attend the lesson without a detailed understanding of the concepts of physics. One of the teachers responded in this regard as:

> *At times students give something meaningless evidences with the subject under discussion. We (teachers) face difficulty to help students because we don't understand why they generate such evidences. (DD)*

This indicates that there are students' who give evidence but difficult to categorize whether theoretical, practical demonstration or from day to day experiences. The sources of such evidence could be from their previous experiences and preconceived knowledge. Such evidence might be categorized under the naïve evidence.

From the discussion above the sources of students' evidence can be categorized as theoretical, practical, and data-based. It was found from the discussion that students generate evidence from physics concepts, laws, and theories they learned before. This is also supported by the literature. Children generate evidence by considering the theoretical relationship between concepts, laws, and theories (Kuhn, 2005; Kuhn & Dean, 2005; Zimmerman, 2000). Practical tasks such as experimental activities, day to day activities, and demonstrations are also indicated by researches as sources of evidence (Zimmerman, 2000). In this study, it was found that there are students who provide evidence from events that occur in day-to-day practices. Students generate evidence from day to day experiences such as rolling ball, playing with mud, moving vehicles, moving objects, flying objects. Data can also be sources of evidence. They generate evidence from values given in the tables, graphs, and recorded data in practical classes.

4.1.2.4 Drawing conclusions

4.1.2.4.1 Scientific view

Most students fail to arrive at valid conclusions because of the abstract nature (if not able to contextualize) of the concepts of the subject. Most of the conclusions are characterized by superstition, biased, and not evidence-based. Students need the concepts to be practically demonstrated, experimented with, and related with day-to-day activities to arrive at a valid conclusion. After providing such assistance students might do things by themselves and arrive at a scientific conclusion. Primary school teachers believed that students have abilities of drawing conclusions about science concepts if the concepts and activities were presented in a clear and authentic way. One of the participants of the FGD, DB, shared one of his experiences about the students' ability to draw conclusions as:

> *When I asked students view about the sayings 'without light there is no life', a particular student responded as 'plants produce their own food by a process of photosynthesis. The process of photosynthesis requires light. We eat the food produced by plants directly or indirectly. Therefore, if there is no light, there is no life'. This shows that students have ability to draw conclusion. (DB)*

4.1.2.4.2 Mixed view

Students might come into science class with some inconsistent views with scientific knowledge. But after physics is introduced they start to recognize that physics is not an illusion rather about the law of natural phenomena around them. But the theoretical conception about physics is not satisfactory to persuade them that physics can be materialized in their context. Therefore, they seek practical evidence to establish these theoretically introduced conceptions. After having ample evidence, students arrive at a conclusion that physics concepts are not an illusion as expected before but can be practically put into practice in a day to day situations. This was summarized by MK3 as:

> *When students come into class they consider physics as not only a very difficult subject but also an illusion. But when they start learning physics objectives, its uses, the relationship with day to day activities, and practical uses they change their mind about physics. They start seeking practical evidences to check whether the concepts they learn from physics laws and theories have any relationship with every day experiences. However, most of the students' conclusions are not accurate enough with the scientific one. (MK3)*

This suggests that students try to draw conclusions using physics concepts, laws, and theories but they fail to arrive at a scientific conclusion.

4.1.2.4.3 Naive view

Primary school teachers believed that only a limited number of students able to draw conclusions scientifically. It was also found from the discussion that, students frequently draw fallacious conclusions, which do not agree with scientific knowledge. Some students consider challenging or opposing others' ideas as evil or ethically wrong. This might be related to respecting someone. It might be considered as a good cultural value for it was told by seniors (Solomon, 2008). There are also times no student is willing to draw a conclusion from evidence if they are not confident enough about their evidence. The challenge with this level of reasoning, as recognized from the discussion, is a lack of conceptual understanding and proper evidence for the ideas of the discussion. This explained by one participant, MB, as:

> *A few students' draw conclusion based on what they learn, see and observe in their own environment. But most students' conclusion is not based on the scientific explanations and evidences. They may draw conclusion based on only ideas they have in mind, or which give sense to them. (MB)*

This shows that such students draw a conclusion based upon what they hear from the teachers, read from textbooks, and their perceptions. For such students providing explanations by itself is the end product or the truth of the matter.

4.1.3 Development of construct map

This section presents the development of the construct map, which was emerged from the review of the literature and the qualitative analysis of students' interviews and teachers' FGD. In this dissertation, the construct is scientific reasoning. Hence, the first part of this section narrates the generation of item map of scientific reasoning construct based on the theoretical frameworks discussed in Chapter 2. It will be followed by a discussion of a person map, based on students' interviews and teachers' FGD, to illustrate the nature of students' scientific reasoning. Eventually, the person-item map will be thoroughly described in detail.

4.1.3.1 Development of item map

The item map comprises four levels of students' reasoning ability which extends from the low ability (factual knowledge) to high ability (drawing conclusion). The levels of the item map depicted in Table 4.10 were obtained from the literature. See Section 2.4 for the detail of how these levels were emerged from the literature.

The level of *factual knowledge* for each style of reasoning involves reading data from a table or figure, identifying units of measurement, formulae, and variables, and stating laws. The level of *explanation of factual knowledge* involves providing variable relationships, providing explanations for the factual knowledge, providing meaning for a given data, and solving physics problems. These two levels represent content knowledge because content knowledge involves knowledge of the facts, concepts, ideas, and theories about the natural world and the explanations (OECD, 2016). *Evidence-based reasoning* involves generating scientific evidence for the explanation provided under level two and identifying valid and invalid evidence; this is categorized under procedural knowledge because procedural knowledge is related to generating scientific evidence to accept or reject a claim (Gott et al., 2008). *Drawing conclusion* involves providing inferences and conclusions based on scientific evidence and this could be categorized under epistemic knowledge for epistemic knowledge involves evidence, hypotheses, models, and theories to establish scientific knowledge (OECD, 2016).

Therefore, the item map depicted in

Table 4.10 demonstrates the construct of scientific reasoning that includes four levels (*Factual knowledge, explanation of factual knowledge, evidence-based reasoning and drawing conclusion*) in a way that incorporates the three forms of scientific knowledge.

Table 4.10: Progresses of Reasoning in terms of Styles of Scientific Reasoning and proficiency levels

		Styles of Scientific Reasoning			**Items**
		Mathematical deductive reasoning	**Model-based reasoning**	**Data-based reasoning**	
Levels of proficiency	**Drawing conclusion**	Draw mathematical model based on the relationship between variables	Draw conclusion using Knowledge of scientific concepts, laws, theories	Provide inferences based the relationship between data	1.4, 2.4, 3.4, 4.4, 5.4, 6.4, 7.4, 8.4, 9.4, 10.4, and 11.4
	Evidence-based reasoning	Generate evidence based on the relationship between variables	Generate scientific evidence from physics laws and theories to support explanations	Generate evidence to strengthen the explanation based on the relationship between data	1.3, 2.3, 3.3, 4.3, 5.3, 6.3, 7.3, 8.3, 9.3, 10.3, and 11.3
	Explanations of factual knowledge	Provide variable relationship among the physical quantities. Use mathematical equations to solve a physics problems	Provide relationship between concepts to provide scientific explanations	Interpret and give meaning for measurement and observation based on the given data	1.2, 2.2, 3.2, 4.2, 5.2, 6.2, 7.2, 8.2, 9.2, 10.2, and 11.2
	Factual knowledge	Recall formulae and units of measurement,	State laws and theories. Recall definitions, constants	Read data from tables and graphs, tell unit of measurements	1.1, 2.1, 3.1, 4.1, 5.1, 6.1, 7.1, 8.1, 9.1, 10.1, and 11.1

Items were developed to represent the levels of scientific reasoning progress indicated in the construct map (see Table 4.10). The final items were developed to represent the levels identified in Table 4.10in a way that model-based reasoning comprises of items (1.1, 1.2, 1.3, 1.4, 2.1, 2.2, 2.3, 2.4, 3.1, 3.2, 3.3, 3.4, 5.1, 5.2, 5.3, & 5.4), mathematical deductive reasoning comprises of items (9.1, 9.2, 9.3, 9.4, 10.1, 10.2, 10.3, 10.4, 11.1, 11.2, 11.3 & 11.4), and data-based reasoning comprises of items (4.1, 4.2, 4.3, 4.4, 6.1, 6.2, 6.3, 6.4, 7.1, 7.2, 7.3, 7.4, 8.1, 8.2, 8.2, 8.3 & 8.4). And also note that items with .1 represent the first level of reasoning, items with .2 represents level 2, items with .3 represents level 3, and items with .4 represent level 4 reasoning.

4.1.3.2 Development of person map

Based on the data collected from the interview of students and videotaped FGD of teachers, the qualitative analysis used thematic analysis. The thematic analysis of students' interviews had depicted that students had demonstrated three different views, namely *naïve view, mixed view, and scientific view* for each of the four level of scientific reasoning, namely *factual knowledge, explanation of factual knowledge, evidence-based reasoning, and drawing conclusion*. During the coding process, students' utterance was coded as naïve when the utterance is intuitive and wrong scientific conceptions. The utterance was coded as a scientific view when students' responses were correct scientific conceptions whereas any utterance that combined the naïve view with the scientific view was coded as a mixed view.

This result was also corroborated by the findings of the qualitative analysis of videotaped FGD, which had focused on teachers' experiences about the nature and characters of their students' scientific reasoning (See Table 4.11). The FGD with school teachers represents hermeneutic phenomena. Teachers revealed various experiences in relation to their students' scientific thinking and reasoning. During the FGD, teachers were encouraged to express their experiences no matter how divergent and contradictory their experiences were. The focus of FGD was not to reach a consensus. In line with this, the researcher had attempted to create a dissensus to address a wide range of teachers' experiences. Thus, hermeneutics guided the role of the researcher as a facilitator of the FGD as well as dissensus inciter.

Table 4.11 An example that depicts how valid is the students' reasoning pattern

Levels of reasoning	Themes emerged from data	Indicative students' responses	Indicative teachers' responses
Factual knowledge	Scientific view	...I first try to remember definitions and formulae to develop better understanding (student 6)	Students' try to recall basic concepts before providing their explanations and solving problems... (teacher DB)
	Mixed view	-	
	Naïve view	..9.8 m/s^2 is truth value... (Student 1)	When I ask questions students try to open exercise books or text book to answer the questions thinking that the facts in the text book are truth or always correct answers (teacher MK3)
Explanation of factual knowledge	Scientific view	Velocity......is the distance covered from starting to end point in a specific direction... (student 12)	After students are exposed to physics concepts they start looking for pattern, create link between known and unknown to construct their own view (teacher MUK3)
	Mixed view	... velocity represents how fast a body moves (student 6)	Students understand the concept of velocity by relating distance and time but they fail to differentiate between speed and velocity.
	Naïve view	F = ma is law of nature discovered by scientists (student 5)	Students believed that there is perfect variable relationship (teacher MK3)
Evidence-based reasoning	Scientific view	The evidence for the presence of gravity is falling objects (student 2)	Students provide evidence from physics concepts, laws and theories (teacher TK4).
	Mixed view	The sources of evidences are teachers, experienced persons and natural phenomenon... (student 7)	Evidence for students is something which can be seen practically, to be demonstrated, and works in daily practices (teacher MUK3).
	Naïve view	...physics laws and theories are true evidence	Students believe that physics concepts, laws and theories are absolute realities (teacher DD)
Drawing conclusion	Scientific view	....F = ma can be improved (student 1)	…an object is said to be in motion when we see that object in reference to other object (teacher DB)
	Mixed view	....based on Newton's first law of motion plants are at rest and will remain at rest for ever (conclusion)	...try to draw conclusion using physics concepts but they fail to arrive on a scientific conclusion (teacher MK3)
	Naïve view	......findings are revealed truth (student 11)	...draw conclusion based on the commonsense knowledge (teacher DD).

The interview and FGD data were transcribed, read thoroughly using relevant documents again and again, and collected memos and explanations. On top of this, Ph.D.

candidates in Physics Education from Addis Ababa University were involved in auditing the data analysis to ensure the credibility of the result. The summary of the qualitative result is provided in a person map demonstrated in Table 4.12 that involves a comprehensive account of the level of grade eight students' views about the four levels of scientific reasoning skills. Students based on their views about the four-levels of scientific reasoning can be categorized into three strata, from Level 1 (naïve view), Level 2 (mixed view), to Level 3 (scientific view). The levels in Table 4.12 represents students' view of each level of scientific reasoning. For example, the first level represents students with naïve views for most of the levels of scientific reasoning and the third level represents students who hold scientific views for most of the levels of scientific reasoning.

Table 4.12: The summary of students' views (person map) towards the four of level of scientific reasoning

Level	Description
3	Student has shown scientific reasoning ability. ➢ Stronger scientific view is observed. ➢ Physics facts, explanations, laws, and theories are temporarily constructed and they are subject to change and modifications ➢ This need to be understood with valid evidence
2	Student has shown mixed scientific reasoning ability ➢ Focused on understanding the product of science by ignoring the process that produced scientific knowledge ➢ Physics facts, concepts, theories and laws are constructed but they are to be remembered, and grasped. ➢ Students focused on recalling of facts and rote memory rather than providing explanations about physical phenomena and relations between concepts of physics
1	Student has shown naïve view towards each level scientific reasoning in Physics. ➢ The reasoning is intuitive and dominated by unscientific conception of the natural/physical world ➢ physical theories and laws are considered to be immutable. ➢ physics facts, concepts, theories and laws are stable, fixed, and certain.

4.1.3.3 Person-item map of the scientific reasoning construct

Based on the literature review, the four levels of reasoning were identified (see Section 4.1.3.1) and students' views were explored towards each level of scientific reasoning (see Section 4.1.3.2). This is then combined to produce the item-person map demonstrated in Figure 4.1. The left-hand side of the item-person map represents students ability (views) towards each level of scientific reasoning and the right-hand side of the map represents the hypothesized construct (scientific reasoning) in terms of the four levels.

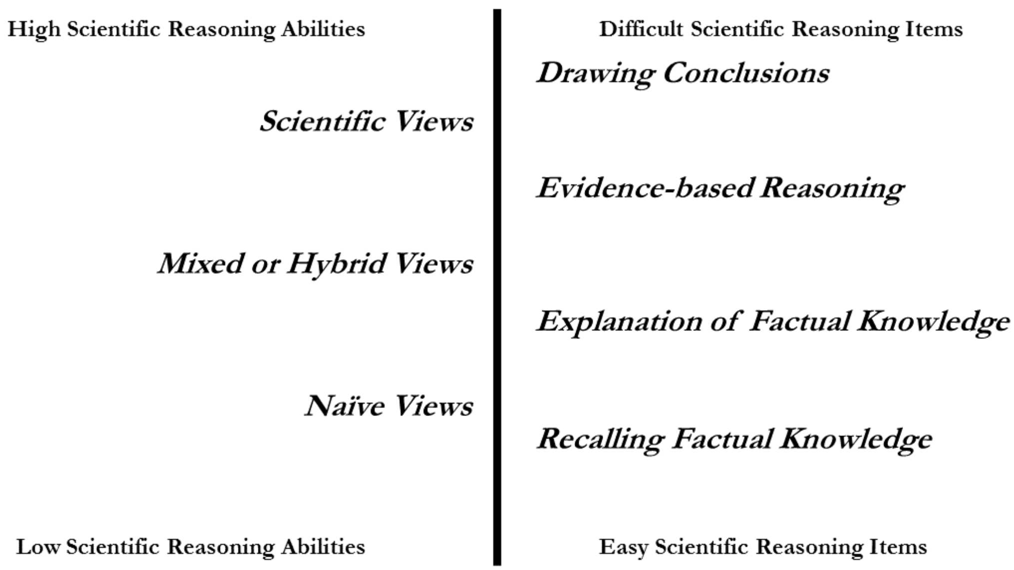


Figure 4.1 The person-item map of SRPT in a continuum.

Sections 4.1.1, 4.1.2, and 4.1.3 dealt with qualitative analysis and findings. It was found that three themes emerged from the students' interview and this was also found to be valid as per the findings of the FGD with primary school teachers (see Table 4.11). The result was considered in the development of items and the construction of the distracters of the items. The distracters of each item were developed in a way that consists of a dominantly naïve view, mixed but towards the naive view, mixed but towards the scientific view, and dominantly scientific view.

The items were then being given to the teachers who were participated in the FGD. According to the teachers' comments, 5 items were rejected and replaced by 5 new items and, 22 items were modified. Moreover, school teachers and a panel of experts had participated in the review process of SRPT. In the sections below, additional analysis will be presented. Qualitative data were collected from both school teachers and a panel of experts iteratively about SRPT items.

4.2 Review of primary school physics teachers

To establish the content and construct validity of SRPT, primary school teachers (who were also participated in the FGD) were participated. These teachers had between ten to 27 years of experience teaching physics at primary schools, all hold the first degree in physics, were participated in the training conducted by the research project TPSS. The themes of the training were teaching scientific reasoning and argumentation, developing activities for dialogical teaching, and how to implement dialogical teaching in the physics classroom.

The teachers' contribution included the development of new items and the modification of existing items. The new items were developed for the final version of SRPT to replace the items that lacked to represent the construct scientific reasoning in terms of difficulty level i.e. lack of contextualization into the students' ability level. In this regard, five new items were developed to represent the initial items. In addition to this, 25 items were modified based on language problems and lack of clarity. In the next sub-sections, some examples will be presented to illustrate the item revision by the teachers.

4.2.1 Exemplars on difficulty level

Three teachers from the six teachers provided an objection about the suitability of question 2 (which consists of items 2.1, 2.2, 2.3, and 2.4). The teachers commented that the items' difficulty level was beyond the Ethiopian curriculum framework because students are required to have a high level of conceptual understanding to answer the questions and Newton's law of motion was presented implicitly. To this end, one of the participant teachers presented alternative items that can suit students' ability levels. He developed items 2.1 and 2.2 that can represent Newton's law of motion explicitly. The items were accepted with minor revision in terms of language and sentence construction (see Figure 4.2).

Initial SRPT Item	**New SRPT Item**
Question 2: Consider a football is placed on a floor covering and a girl pulled forward the floor covering. 2.1 As she pulls the covering forward which way do you think the ball will probably move? a) Forward b) Backward c) No motion of the ball d) I don't know which way will it move, unless I observe the phenomena	**Question 2:** Suppose a ball is moving in a smooth horizontal frictionless surface. 2.1 Which physics law states that <u>if the ball is in motion with constant velocity it will continue its motion if no net external force is exerted</u>? a) Newton's 1st law b) Newton's 2nd law c) Newton's 3rd law d) There is no physics law related to the given statement.
2.2 Why do you select the above answer for question number 2.1? It is because a) The ball will move forward together with the covering b) The tendency of the ball is to remain at rest, c) Only the floor covering is moving d) The information given is not satisfactory	2.2 What will happen to the motion of the ball <u>**if no net external force is exerted**</u> in question 2.1? a) It will slowdown and comes to rest. b) It will continue in motion with constant velocity a) It will speed up b) It is difficult to predict

Figure 4.2 Initial SRPT item, that were identified under the category of 'difficult level' and the modified SRPT item based on the recommendations of the teachers

The change made on items 2.1 and 2.2 had forced us to develop new items that can replace the initial items 2.3 and 2.4 of SRPT. Therefore, new items that can represent items 2.3 and 2.4 were developed, for the final version of SRPT, by the researcher to represent the level of evidence-based reasoning and drawing conclusion.

4.2.2 Exemplars on lack of clarity

In terms of clarity and language problems, 25 items were modified based on the teachers' comments and suggestions. Two items were presented as below as exemplars.

One of the participating teachers commented on the clarity of question 10 saying that it is confusing for primary students to identify the two scenarios described in the question. He suggested the cases to be expressed explicitly for the students to understand the question easily. In addition to this, one of the teachers identified problems related to the answer options of the initial version of SRPT. All options could be the correct answers for the item.

For this reason, the question was re-written, the cases were identified clearly by adding 'case 1' and case 'case 2'in sentence construction, and answer options were re-written (see Figure 4.3).

Initial SRPT Item	**Revised SRPT Item**
Question10: A man lifts vertically 30kg object to a height of 2m, and 20kg mass vertically to a height of 3m. 10.1 Which of the following variables *is* required to calculate gravitational potential energy? a) mass and height b) mass and gravitational acceleration. c) height and gravitational acceleration d) mass, height and gravitational acceleration	**Question**10: A man lifts an object of 30 kg vertically to a height of 2 m (case 1), and an object of 20 kg vertically to a height of 3 m (case 2). 10.1 Which of the following variables *are* required to calculate gravitational potential energy? a) Mass and velocity b) Height and velocity c) Velocity and gravitational acceleration. d) Mass and height
10.2 In which case the man does requires more work? a) In the first case b) In the second case c) The same work for both cases	10.2 In which cases work done by the man to lift the object is greater? a) In the first case b) The same work for both cases c) In the second case

d) Impossible to predict	d) Impossible to predict

Figure 4.3 Initial SRPT items, that were identified under the category of 'lack of clarity' and the modified SRPT item based on the recommendations of the teachers

To add one more exemplar, teachers raised some issues related to the construction of the sentence structure on item 10.2 to make the item clear. According to one of the primary school teachers, the item was unclear for the students. Careful inspection has shown that the writing had a problem related to grammatical arrangement. Therefore, the item was re-written to make it clear for the students as shown in Figure 4.3. In the same way, the other 23 items were also modified according to the comments, suggestions, and recommendations of the primary school teachers.

In addition to the primary school physics teachers, the SRPT was also reviewed critically. In the following sub-sections, some exemplars of the review of pannel of experts will be presented.

4.3 Review of panel of experts

To establish content validity, the panel of experts had iteratively reviewed the SRPT test items. The panel of experts was composed of two Ph.D. candidates in Physics Education at Addis Ababa University. Both candidates had taken a course, which was entitled as *Scientific Reasoning and Argumentation,* during their doctoral study.

In addition to this, each candidate had more than ten years of experience in teaching physics to pre-service teachers. Hence, these candidates had the relevant expertise to review SRPT items. First, the panel reviewed the initial SRPT items and identified items that had issues related to the content validity. Out of the initially prepared 44 SRPT items, the panel designated nine items as having problems linked with its content. By and large, the nine items, with content validity issues, can be categorized into two as those with vague wordings and those that need further improvisations and modifications due to concerns in scientific reasoning contents. Six SRPT items were labeled as *vague wording* while the remaining three SRPT items as *scientific reasoning contents*. In the next two sub-sections, explanations about each category will be briefly presented. Exemplars from both categories will be provided to illustrate particular validity issues to substantiate the analysis in this phase.

Under the category of *vague wording*, items that were imprecise in their wording, problematic to smoothly understand the question, and had vague vocabulary were identified. Under the category of *scientific reasoning contents*, items that were not relevant to measure the scientific reasoning levels of grade eight students, items that had inaccurate physical concepts and beyond the level of grade 8 physics of Ethiopian physics curriculum, and were not contextual to the students cultural settings were selected.

4.3.1 Exemplars on *vague wording*

The panel identified item 8.4 of SRPT as having vague wording, and hence have validity issues. Figure 4.4 presents this item as an example of the SRPT item under this category. The item was designed to probe students level four (drawing conclusions) scientific reasoning.

Initial SRPT Item	Revised SRPT Item
8.4 What would you conclude if students use 7.5 volt battery in the above experiment?a) The value of the current measured by the students will vary in the range between 2.95 and 3.05A. b) The value of current measured by the students will be greater than 2.95 A c) The value of current measured by the students will be greater than 3.05A d) The measured value of current will be greater than 2.95A and 3.05A	8.4 What would you conclude if students use 7.5 volt battery in the above experiment? a) Student A and Student B will measure current that lies in the range between 2.95 and 3.05A. b) Student A and Student B will measure current less than 2.95 A c) Student A and Student B will measure current less than 3.05A d) Student A and Student B will measure current greater than 2.95A and 3.05A respectively.
1.2 How can you describe your state during this journey from Addis to Hosanna? a) I am in motion b) I am in motion and also at rest c) I am at rest d) Neither at rest nor in motion	1.2 How can you describe your state of motion during this journey from Addis to Hosanna? a) I am always in motion b) I am some times in motion and sometimes at rest c) I am always at rest d) I am neither at rest nor in motion

Figure 4.4 Initial SRPT item, that were identified under the category of 'vague wording' and the modified SRPT item based on the recommendations of the panel of experts

Both Ph.D. candidates, who are serving as a panel of experts to this study, had designated the abovementioned SRPT items as having unclear wordings. The panel assumed that imprecise phrasing observed in the multiple choices (particularly on the wording of choice d and the phrase “by the students”), hence, the panel concluded that lack of clarity might lead students to frustration and anxiety. One of the members of the panel, for example, suggested that the phrase “by the students” should be modified to show clearly that there are two students: student A and Student B – who are doing the experiment. And the panel suggested considering rephrasing to show that there are two students and to provide more clarity to the multiple choices of this SRPT item. Therefore, this SRPT item was edited to include the suggestions of the panel of experts and to add to its content validation. See Figure 4.3 for the initial and modified SRPT item 8.4.

To add more exemplars, another item of SRPT (item 1.2) was evaluated as lacking precision in wording. The initial item 1.2 (please see Figure 4.4) has a phrase "…describe your state during..."One of the experts said that "the phrase in this question, which is …describe your state during...might have multiple meanings. What state? I suggested saying…describe your state of motion during..."In addition to pointing the lack of word clarity of this item, the comment of the panel had given insights about how some of the students might comprehend the question. In line with the feedback of the panel, item 1.2 was modified to avoid further confusion and to increase the overall content validity of the SRPT test items. See Figure 4.3 for the initial and modified SRPT item 1.2.

4.3.2 Exemplars on scientific reasoning contents

In this section, exemplar items that depict content validity issues on the ground of the content of the scientific reasoning represented in the item will be discussed. Item 9.4 is attempting to probe students' ability to draw conclusions. In this study's framework, drawing conclusions is considered as the highest level of reasoning expected to be observed on grade eight students.

Item 9.4, which is shown in Figure 4.5, was identified by the panel as having problems in relation to measure that specific level of reasoning. Hence, it was considered as having content validity related issues that was required to be addressed. One of the experts said, "I don't see the capacity of this item to probe the "drawing conclusion" level of reasoning. For students to draw sensible conclusions, students should show a chain of reasoning from the observation phases to the conclusion stage. Item 9.4 lacks to depict such assessment trends."

This insight from the panel led to the modification of this item. The item was edited in both its questions and multiple choices. By changing the multiple choices into graphical representations, item 9.4 had provided an opportunity for students to display a chain of reasoning that goes from determining the relation between force, acceleration, and mass to inferring visual representation of the relationship between force, acceleration, and mass. Figure 4.5 provides both the initial item of 9.4 and the modified item of 9.4 side by side. In brief, this figure depicts how the obtained feedback was considered and used to change the item.

<table>
<tr><th>Initial SRPT Item</th><th>Revised SRPT Item</th></tr>
<tr><td>9.4 Which one of the following equations represents the best variable relationship in question 9? (Where k is constant; x and y are variables)

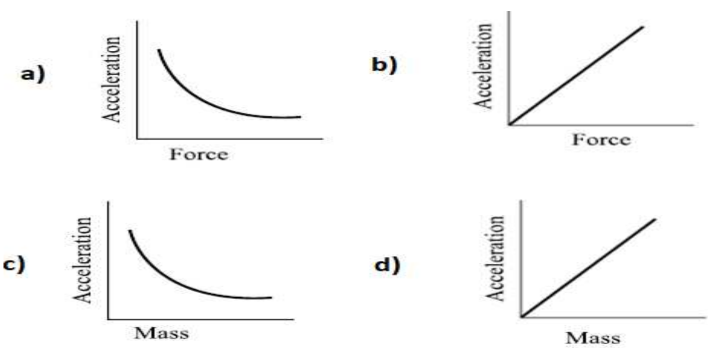
a) $y = kx$
b) $y = kx^2$
c) $y = k/x$
d) $y = k/x^2$</td><td>9.4 Which one of the following graphs represents your answer to question number 9.3?

</td></tr>
</table>

Figure 4.5 Initial SRPT item, that were identified under the category of 'vague wording' and the modified SRPT item based on the recommendations of the panel of experts

In addition to these two exemplars, the panel of experts had suggested providing directions that are clear and easily understandable by grade eight students. Besides, the same comment was forwarded about questions in general. The panel advised using clear language in both directions and questions that can be easily understandable by the targeted test-takers population. During the revision of SRPT items, this study had considered experiences and insights gained from students' interviews, teachers' focus group discussion, and experts review.

To sum up, in this chapter students' views about the four levels of scientific reasoning skills explored. the four levels of reasoning that were identified from the review of literature were also revealed using students' interviews. Only 8% of students had shown a scientific view for the fourth level, 16% had shown a scientific view for the third level, 25% of students had shown a scientific view for the second level, and 33% had shown a scientific view for the first level. Based on the four levels of the scientific reasoning and three views towards each level of scientific reasoning, the item-person map was developed as shown in Figure 4.1, and this, in turn, guided the development of SRPT. Furthermore, the content validity of SRPT was established using teachers' review and a panel of experts review (seeFigure 4.2, Figure 4.3, Figure 4.4, and Figure 4.5to overview exemplar SRPT items reviewing and revising processes). Out of 44 items reviewed by the teachers, 5 items were replaced by new items, and 22 items were modified. Out of 44 SRPT items reviewed by the panel of experts, nine items (six as having vague wording while three as having scientific reasoning content) were identified and revised to increase the content validity of SRPT. The panel had participated in identifying items that lack validity to the test purpose and in providing insightful feedback during the revision process. The revised version of SRPT is provided in Appendix A.

In the next chapter, the quantitative validity and reliability of SRPT will be presented. The Rasch model had been used for the psychometric analysis of both validity and reliability. A corpus of data for the quantitative phase was collected during the filed-test phase. The field test had used all 44 SRPT items to investigate validity and reliability quantitatively. Two hundred and three in the first testing and two hundred and forty-two grade eight students from six schools in Hadiya zone had taken the revised SRPT.

Chapter 5 Quantitative analysis and Results

In this chapter, the quantitative analysis of the pilot test and the field test will be discussed. In this quantitative stage of analysis, the psychometric properties of SRPT will be measured using Rasch analysis. This chapter also presents the Rasch analysis of SRPT that include reliability, item discrimination ability, item difficulty level, various fit analysis, and Wright map analysis. These analyses were utilized to validate SRPT and to select the final version of SRPT items.

5.1 Result of pilot testing

5.1.1 Demographic characteristics of the pilot test participants

The pilot test was conducted in Kembata Zone, SNNPR. Kembata zone was selected for the pilot testing due to the textbooks and curricular materials are similar to Hadiya Zone. In addition to this, all the schools and teachers who participated in the pilot test did not participate in the TPSS project. The pilot testing was administered by the school teachers. The researcher had participated in facilitating the testing of grade 8 students. In the pilot test, 203 grade eight students from six schools had participated.

After making the necessary revisions based on the suggestions of the panel of experts and schools during the qualitative content and construct validation stages, the pilot test was administered to a cohort of 203 grade eight students. Before the commencement of the SRPT test, students were informed about the purpose of the SRPT test. In addition, general directions were also communicated. Six school physics teachers invigilated the test. The researcher took notes of various issues students had raised when taking the SRPT. Moreover, the time taken to finish the test by each student had been recorded. Initially, the allotted time to take SRPT was 1 hour. However, most of the students had finished the test in one hour and twenty minutes. Thus, the allocated time was revised. Such nuances of data and the quantitative Rasch analysis were used to establish the construct validity as well as the overall quality of the test. In what follows, the quantitative Rasch analysis will be presented. Winsteps 3.68.0 software had been used for data analysis purpose (Linacre, 2020).

In this dissertation, to establish the construct validity and reliability of the SRPT, the Rasch model was used. To measure various psychometric properties of SRPT, the following common Rasch analysis was done. These include unidimensionality, item and person reliabilities, fit statistics, bubble charts, and Wright map analysis. The result of these analyses was used to revise items of SRPT before the field testing stage.

Before the item creation, the person-item map, which consists of four levels, and three students' views were established (see Figure 4.1). Following this, the items that represent each level were developed. To determine whether the items properly represented the intended levels in the construct map, fit statics along with other indicators were used. Fit statistics provide the discrepancies between the expected Rasch modeling and actual data of the test. It helps to identify misfit items along with the possible causes. The probable reason for items to be poor to measure the wanted construct could be using unclear words, conceptual problems, the ambiguity of the questions, and measuring more than one construct (Bond & Fox, 2007; Engelhard Jr, 2013). The Rasch model provides infit and outfit indicators with values measured in terms of MNSQ and ZSTD.

5.1.2 Fit Statistics of SRPT

To determine the degree to which the data fit the Rasch model, item Infit and Outfit statistics were calculated. Infit statistics deals generally with the pattern of responses in terms of theoretical prediction. It evaluates whether the theoretical prediction agrees with the data (students' response). Outfit statistics deals with simple responses of students concerning item difficulty. It is concerned about determining whether students' ability aligns with item difficulty. This means that it evaluates whether high achieving students answered easy items and vice versa. The recommended range of infit and outfit for such a kind of multiple-choice tests is 0.7 to 1.3. The values greater than the interval indicates the data are less predictable to the model and the values less than the recommended interval indicates the data are more predictable to the model. MNSQ and ZSTD are a means to provide the results of Infit and Outfit values. The acceptable range of ZSTD value is between -2 and 2 but only important when the MNSQ values are not in the acceptable range (Bond & Fox, 2015; Linacre, 2020).

For the pilot test, the Infit MNSQ values of SRPT fell in the range of .88 to 1.11 (see Table 5.1) which indicates that the learners' reasoning abilities match the item difficulties according to the reasoning levels established in Table 2.5. The Outfit MNSQ values fell in the range between .88 and 1.24 (except item 8.3) which also indicates that the responses have met the expected patterns (see Table 5.1). The ZTSD values for the items fell in the accepted interval (-2 to +2) except for items 8.3, 11.3, 5.3, 5.2, and 3.3. This also indicates that students' responses to the items in terms of item difficulty are in an expected manner except for the five items.

The validity of a scale becomes a threat when there are more misfit (underfitting) items with unacceptable values for MNSQ, and placed at the right side of the bubble chart (Linacre, 2020; Martin & Jamieson-Proctor, 2019). The items which have an unacceptable range for MNSQ value should be inspected seriously. However, items with unacceptable ZSTD values may not cause a threat to the validity of a test if they are in the acceptable range of MNSQ value. In this regard, item 8.3 and 11.3 are subject to modifications or rejection. Items 5.3, 5.2, and 3.3 were not subjected to revision or rejection, on the ground of what is observed in the bubble chart. The justification is presented in detail in Section 5.1.3.

Table 5.1 Fit statistics for the first version of SRPT

Item No	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	Pt. Corr.
Item 1.1	.98	-.3	.94	-.6	.29
Item 1.2	.96	-1.1	.95	-1.2	.34
Item 1.3	1.10	1.0	1.23	1.8	.00
Item 1.4	1.03	.7	1.06	1.1	.19
Item 2.1	.96	-1.2	.96	-.9	.33
Item 2.2	.98	-.5	.98	-.4	.29
Item 2.3	1.08	1.5	1.11	1.7	.11
Item 2.4	1.01	.1	1.04	.4	.20
Item 3.1	.99	-.1	1.00	.1	.25
Item 3.2	.92	-1.8	.91	-1.6	.40
Item 3.3	.88	-3.2	.88	-2.8	.46
Item 3.4	.94	-1.6	.93	-1.4	.37
Item 4.1	.99	-.2	.98	-.3	.28
Item 4.2	.95	-1.4	.94	-1.4	.36
Item 4.3	.92	-1.7	.91	-1.8	.40
Item 4.4	.99	-.2	.99	-.1	.27
Item 5.1	1.00	.3	1.00	.1	.25
Item 5.2	.86	-3.4	.86	-3.3	.48
Item 5.3	.88	-3.0	.88	-2.9	.45
Item 5.4	.89	.1	1.02	.2	.22
Item 6.1	1.06	.5	1.20	1.2	.04
Item 6.2	1.09	1.9	1.12	2.1	.10
Item 6.3	1.03	.3	1.17	.9	.07
Item 6.4	1.04	.6	1.05	.7	.18
Item 7.1	1.07	.6	1.20	1.2	.03
Item 7.2	.95	-1.3	.94	-1.5	.35
Item 7.3	.95	-1.0	.93	-1.2	.34
Item 7.4	1.09	2.1	1.08	1.6	.12
Item 8.1	1.01	.1	1.04	.3	.19
Item 8.2	.97	-.8	.97	-.7	.31
Item 8.3	1.13	.9	1.45	2.3	-.13
Item 8.4	.99	-.2	.98	-.3	.27
Item 9.1	.95	-.8	.92	-.9	.34
Item 9.2	1.04	.9	1.04	1.0	.20
Item 9.3	1.06	1.5	1.05	1.2	.16
Item 9.4	.99	-.2	.99	-.1	.27
Item 10.1	1.00	.1	1.00	-.1	.26
Item 10.2	.95	-1.0	.95	-1.1	.34
Item 10.3	1.03	.6	1.03	.6	.22
Item 10.4	1.07	1.1	1.07	1.4	.10
Item 11.1	.98	-.5	.98	-.5	.30
Item 11.2	.98	-.5	.98	-.5	.30
Item 11.3	1.13	2.8	1.11	2.8	.06
Item 11.4	.99	-.1	1.00	-.1	.25

Let us consider two items (item 8.3 and 11.3) to explain how the fit statistics was used to identify item problems, causes, and how to improve the items to measure the predicted levels of scientific reasoning. As shown in Table 5.1 the outfit MNSQ value of item 8.3 is 1.45 and outfit ZSTD is 2.3 which is not in the acceptable range. This indicates that the students were not answered the item according to their ability level (see the explanation for outfit value above). This means that less able students were answered better the item than most able students. It was, initially, developed to measure students' ability of evidence generation, and the subsquent item 8.4 was developed to measure students' ability of drawing conclusion. However, based on the students' responses item 8.3 appeared to be more difficult than item 8.4 which contradicts theoretical prediction. The reason for this mismatch was found to be the way item 8.3 was constructed to represent the level of evidence-based reasoning. the careful inspection had shown that the initial version of item 8.3 tends to be more conclusive than demanding evidence from the data for the explanation provided at level 2. For this reason, item 8.3 was re-written in a way that can request students to generate evidence for the explanation provided at level 2 based on the data given in the table. In the same way, item 11.3 has an infit ZSTD of 2.8 and outfit ZSTD of 2.8. These values might also indicate that the item is not measuring the students' ability as predicted in the construct map (see the explanations for infit value above). The item was intended to be answered by considering item 11.2. However, careful inspection suggests that the item has a lack of coherence and consistency with the alternatives of item 11.2; initially, they were supposed to be related to giving a meaningful pattern. Thus necessary revisions were done for this item 11.3, to make it coherent and consistently align with 11.2.

5.1.3 Bubble Chart

Figure 5.1 presents the bubble chart of SRPT using the pilot test data. The validity of a scale becomes a threat when there are more misfit (underfitting) items on the right-hand side of the bubble chart than items (overfitting) on the left-hand side (Martin & Jamieson-Proctor, 2019). As indicated in the bubble chart (see Figure 5.1), both items (item 8.3 and 11.3) are placed outside of the expected range (+2 and -2). In addition to this, item

8.3 correlates negatively with other items; item 11.3 has a very low correlation with other items to measure the level of the predicted construct (see Table 5.1). This means that they are not contributing to measure a single construct scientific reasoning as established in the construct map. This indicates that the items are probably measuring some other construct than the intended construct i.e. scientific reasoning.

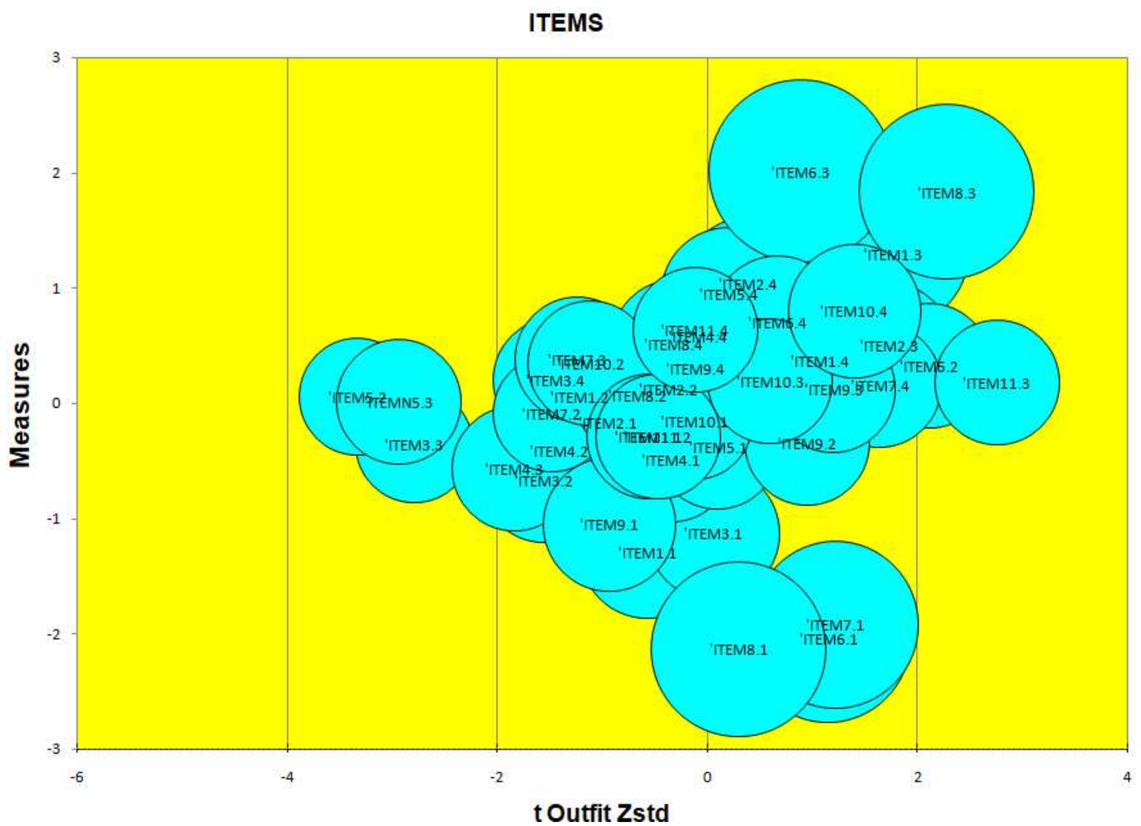


Figure 5.1 Bubble chart of the SRPT shows how each of the 44 items fit the scale

Table 5.2 demonstrates that item 8.3 was replaced by another item that can better suit item 8.3 and properly represent level 3; item 11.3 was revised in a way that can be consistent with the alternatives of item 11.2 to make the item clearer for students.

Table 5.2 Sample item that depicts the initial and revised SRPT items

Initial SRPT Item	**Revised SRPT Item**
8.3 Why did not both students get exactly the same measurement in the table? a) Measurement can't always give exactly the same value b) One of the student is not skilful in measurement c) There is some problem in measuring devices d) Both students are wrong	8.3 Why do you select the above answer for question number 8.2? It is because a)The measurement is taken on voltage and current at constant resistance b) The measurement is taken on current by varying voltage at constant resistance c) The measurement is taken on voltage and resistance is calculated d) The measurement is taken on voltage, current, and resistance.
Question 11 Your neighbour needs a 2.0-ohm resistor but he has only 4.0-ohm resistors. He asks an electrician to help him how to produce a 2.0-ohm resistor. The electrician tells him that to connect two 4.0-ohm resistors in parallel to get a 2.0-ohm resistor. 11.2 Do you agree or disagree with the electrician? a) I agree b) I disagree c) I am not sure	Question 11 Your neighbour needs a 2.0-ohm resistor but he has only 4.0-ohm resistors. He asks an electrician to help him how to produce a 2.0-ohm resistor. The electrician tells him that to connect two 4.0-ohm resistors in parallel to get a 2.0-ohm resistor. 11.2 What is your opinion about the electrician advice? a) I agree with the electrician b) I disagree with the electrician c) I agree to some extent. d) I am not sure
11.3 Why do you agree or disagree with the electrician in 11.2? It is because a) In parallel connection voltage is constant b) The total resistance of two 4.0-ohm resistors in parallel gives a 2.0-ohm resistor c) The total resistance of two 4.0-ohm resistors in series gives a 2.0-ohm resistor d) He is the electrician who gave the advice.	11.3 Why do you agree, disagree, agree to some extent or not sure with the electrician in 11.2? It is because a) In parallel connection voltage is constant. b) The total resistance of two 4.0 -ohm resistors in parallel gives a 2.0 -ohm resistor. c) The total resistance of two 4.0 -ohm resistors in series gives a 2.0-ohm resistor. d) Resistance cannot be added, rather made with fixed values.

On the other hand, the other three items (item 5.3, item 5.2, and item 3.3) have the ZSTD values outside the expected range and are placed on the left side of the bubble chart (see Table 5.1 and Figure 5.1). These items are not a threat for construct validity; the validity of a scale becomes a threat only when there are more misfit items on the right-hand side of the bubble chart. The results above indicated that the first three items are redundant (there are more items to represent the same area of the construct) and not necessarily remained to be part of the tool unless the rejection affects the overall quality of the tool. These items can remain in the tool, rejected, or modified to represent students' reasoning ability properly. They are not a threat to the validity of the tool. However, the items were modified before including in the final version. Item 3.3 was modified in a correct grammatical arrangement and item 5.3 was re-written in a way to be answered by referring to item 5.2.

5.1.4 Wright map analysis of SRPT

Even though the item-person map of SRPT using the pilot test data had demonstrated the good distribution of items across the map, there were gaps among items (see Figure 5.2). For instance, there was a gap between item 7.1 and item 1.1; item 3.2 with item 3.1 and item 9.1; item 1.3 with item 8.3; and item 1.3 with items 2.4, and item 5.4.

The location of item 7.1 was -1.92 (ability measure) and the location of item1.1 was -1.29. This suggests that students with ability between -1.29 and -1.92 may not be well targeted with this scale. This tells that there is a need for additional items to fill this potential gap between the items, modification of the construct map or modification of the three items is required to make the items somehow more difficult. In the same way, the location of item 8.3 on the map was 1.83 and the location of item 1.3 was 1.3. This also indicates that there is a significant gap between the two items. This also tells there is a need to incorporate an additional item to fill the gap and needs reconsideration in the final tool. The gaps between the other items are not worthy to consider.

```
       PERSONS - MAP - ITEMS
            <more>|<rare>
 2                +  ITEM6.3
                  |
                  |  ITEM8.3
                  |T
                  |
               .  |
                  |
                  |  ITEM1.3
                  |
                  |
 1             # T+  ITEM2.4   ITEM5.4
             ###  |S
              ##  |  ITEM10.4
            .###  |  ITEM6.4
               .  |  ITEM11.4  ITEM4.4
          ######  |  ITEM2.3   ITEM8.4
              ## S|  ITEM1.4   ITEM10.2  ITEM7.3
         #######  |  ITEM6.2   ITEM9.4
        ########  |  ITEM10.3  ITEM11.3  ITEM3.4   ITEM7.4
        .#######  |  ITEM1.2   ITEM2.2   ITEM5.2   ITEM8.2   ITEM9.3
 0        ######  +M ITEMN5.3
        ######## M|  ITEM7.2
       .########  |  ITEM10.1  ITEM2.1
         .######  |  ITEM11.1  ITEM11.2  ITEM9.2
           .####  |  ITEM3.3   ITEM4.2   ITEM5.1
           #####  |  ITEM4.1
           .####  |  ITEM4.3
         .###### S|  ITEM3.2
               #  |
             .##  |S
-1           ###  +
             ###  |  ITEM3.1   ITEM9.1
              ## T|
                  |  ITEM1.1
               .  |
               #  |
                  |
                  |T
                  |
                  |  ITEM7.1
-2                +  ITEM6.1
                  |  ITEM8.1
                  |
                  |
                  |
```

Figure 5.2 Item person map for the SRPT: Logit scale with increasing values from bottom to top distributes the person abilities (on the left) and the item difficulties (on the right).

5.1.5 Reliability and Separation Indices of SRPT

Table 5.3 shows the separation indices of SRPT for the pilot test. The item separation index of SRPT was found to be (5.28) and the item reliability was (.97). This indicates that the person sample is large enough to confirm the difficulty of the items' hierarchy to measure the expected levels of reasoning ability (Linacre, 2020). The person reliability was (.63) and the person separation index was (1.31) which indicated that the items are not discriminating properly the high and low achievers (Boone et al., 2013). This suggests that the SRPT needs some revisions to separate the test-takers into high, medium, and low achievers.

Table 5.3: Summary of person and item statistics: reliability coefficients and separation indices

				INFIT		OUTFIT	
	Mean	Separation	Reliability	MNSQ	ZSTD	MNSQ	ZSTD
Items	15.6	5.28	.97	.99	.0	1.02	1.0
Students	94.3	1.31	.63	1.00	-.2	1.02	-.1

5.1.6 Revision of the construct map

Based on the students' responses during the pilot testing of the SRPT, there was no need to change the basics of the initial construct map. However, the responses of students' for some items did not match the theoretical prediction as hypothesized in the construct map. Most items were distributed as predicted except items that are constructed to represent evidence-based reasoning (see item person map above). As shown in the construct map, evidence-based reasoning appeared to be more difficult for students than drawing conclusion which contradicts the theoretical prediction.

To identify the cause of this unexpected finding, this was inspected carefully. The researcher tried to identify whether the quality of items or order of construct map in terms of reasoning level causes such mismatch between theoretical prediction and data from students' responses. It was discovered that the disparity between theory and data originated from the way how items were constructed for the level of evidence-based reasoning. The evidence for this is that items intended to measure the level of evidence-based reasoning were poor and did not fall in the expected interval of Rasch modelling. For example, item 8.3 and

11.3 were not fit items (see the description under Section 5.1.2 and 5.2.2); item 1.3 has zero correlation with the other item to measure the construct. This suggests that the way how the items were constructed lacks alignment with the expected performance. Therefore, this indicates that there is no need for modification of the construct map in terms of the mentioned items.

However, in the pilot test, students' responses to the items of SRPT suggested that there was a gap between items 8.1, 6.1, and 7.1, with the item 1.1. The three items (8.1, 7.1, and 6.1) were constructed to measure the first level of data-based reasoning. This suggested that the first level of data-based reasoning was initially written with easier performance in the construct map and the corresponding items that represent the level than model-based and mathematical reasoning. The first level of the initial item map (see Table 4.10) for model-based reasoning demanded students to *tell units of measurement*; the second level demanded them to *Interpret and give meaning for measurement data*. This suggests that there is a gap between the first and the second level in terms of data-based reasoning. It was discovered from this that, before the students give interpretation and meaning for data, they need to differentiate the value of one variable from the other. To fill this gap and bring consistency with the two styles of scientific reasoning, the construct map was modified as in Table 5.4. However, this brings the possibility of another level of reasoning, particularly below the first level of reasoning (Factual knowledge) to be further investigated in future research endeavors.

Table 5.4 Initial and revised construct map for the first level of scientific reasoning.

Level of reasoning		Factual knowledge
Initial construct map	Mathematical deductive reasoning	Recall formulae
	Model-based reasoning	State laws and theories. Recall definitions, constants
	Data-based reasoning	tell unit of measurements
Revised construct map	Mathematical deductive reasoning	Identify given variable, unknown variables and formulae; differentiate one variable from other variable, one formula from the other.
	Model-based reasoning	Identify physics concepts, laws, and theories; differentiate one statement of law from the other.
	Data-based reasoning	Identify data from tables and graphs; differentiate value of one variable from the other

In the same way, to create consistency of the first level of data-based reasoning with other styles of scientific reasoning (mathematical deductive and model-based reasoning), modifications were also made for the two styles of scientific reasoning as shown in

Table 5.4. This modification was also made for the items that represent the first level of scientific reasoning.

Table 5.5 presents an exemplar of item modification to represent the revised construct map. Item 8.1 was constructed initially to represent the first level of the initial construct map (to identify units of measurement). Based on the modification of first level reasoning the item was modified as Table 5.5 which demanded to identify a given data from a table.

Table 5.5 Exampler item 8.1 to demonstarte the revision process of SRPT items developmnet phase

Initial SRPT Item	**Revised SRPT Item**
Item 8.1 Which of the following units of measurement is used to measure voltage? **a) volt** b) ampere c) ohm d) ammeter	Item 8.1 Which one of the following measurements is **<u>not</u>** value of current measured by the first student? a) 1.40 A b) 0.73 A c) 2.20 A **<u>d) 1.45 A</u>**

Moreover, the initial version of the construct map did not give any clue about students' ability of each level of scientific conclusion. It was written in vague and general terms; it did not demonstrate students' ability for each level of scientific reasoning. For this reason, the initial construct map was modified based on the qualitative and quantitative findings of the two levels. For example, both qualitative and quantitative results had shown that most students failed to generate scientific evidence and draw a scientific conclusion; most students generate invalid evidence and draw conclusion based on intuitive knowledge. Therefore, the initial construct map was modified by including their performance level as shown in Table 5.6.

Table 5.6 Initial and revised construct map for the second level of scientific reasoning

Level of reasoning	Evidence-based reasoning	Drawing conclusion

Initial construct map	Mathematical deductive reasoning	Generate evidence based on the relationship between variables	Draw mathematical model based on the relationship between variables
	Model-based reasoning	Generate scientific evidence from physics laws and theories to support explanations	Draw conclusion using Knowledge of scientific concepts, laws, theories
	Data-based reasoning	Generate evidence to strengthen the explanation based on the relationship between data	Provide inferences based the relationship between data
Revised construct map	Mathematical deductive reasoning	Students attempt to generate evidence based on the relationship between variables explanation but they struggled to generate scientific evidence.	Students try to draw conclusion based on the evidence but most students fail to arrive on scientific conclusion.
	Model-based reasoning	Students attempt to generate evidence but most students faced difficulty to generate with correct physics laws and theories.	Students face difficulty of drawing conclusion based on scientific facts, concepts, laws and theories.
	Data-based reasoning	Students attempt to generate evidences for the explanations but struggled to correctly interpret the meaning of data.	Students face difficulty of drawing conclusion based on the given data.

To sum up, the initial construct map was to some extent revised in a way that can create a better link between each level of reasoning, and in a way that can demonstrate the students' ability based on the results of this study for each level. The initial and final construct maps were depicted in Table 4.10 and Table 5.7 respectively.

Table 5.7: Final version of construct map based on the students' reasoning ability

		Styles of Scientific Reasoning		
		Mathematical deductive reasoning	**Model-based reasoning**	**Experimental reasoning**
Levels of Proficiency	**Drawing conclusion**	Students try to draw conclusion based on the evidence but most students fail to arrive on scientific conclusion. Characterized with naïve conclusion.	Students face difficulty of drawing conclusion based on scientific facts, concepts, laws and theories. Characterized with naïve conclusion.	Students face difficulty of drawing conclusion based on the given data. Characterized with naïve conclusion.
	Evidence-based	Students attempt to generate evidence based on the	Students attempt to generate scientific evidence to support	Students attempt to generate evidence for the explanations of

	reasoning	relationship between variables explanations but they are struggled to generate scientific evidence. Characterized with mixed evidences	explanations but lacks consistency between evidence and factual knowledge. Students' evidence is characterized with day to day experiences but not accurately.	variable relationship but struggle to correctly interpret the meaning of data. Characterized with incorrect variable relationship.
	Explanations of factual knowledge	Most students able to solve physics problems using mathematical equations; provide explanations about variable relationships. It is characterized by creating link between known and unknown.	Most students are able to provide explanations based on the physics concepts, laws and theories. It is characterized by creating link between physics concepts, laws and theories with daily practices	Most students are able to give meaning for measurement values and observation based on the given data in table and graphs**.** Characterized by comparing data to understand meaning of variable relationship.
	Factual knowledge	Students are able to identify given variable, unknown variables and formulae; differentiate one variable from other variable, one formula from the other.	Students are able identify physics concepts, laws, and theories; differentiate one statement of law from the other.	Students are able to read data from tables and graphs; differentiate value of one variable from the other

Generally, the Rasch analysis of the pilot test indicated that SRPT consists of items that are a threat to its construct validity. SRPT items that were prepared to measure level 3 and level 4 were not also positioned hierarchically as per the established construct map (see initial item person map). In addition to this, it has also low person and item separation, low person reliability (see Table 5.3), and a significant gap was observed between items 8.1, 6.1, and 7.1, with item 1.1 (see Figure 5.2). This suggests that the test needs improvement to measure a single construct according to the order of the construct map. In brief, the quantitative validation of the pilot test indicated that SRPT needs further revisions to improve its quality.

After making the necessary revisions to items of SRPT that were identified as problematic using quantitative analysis of Rasch modelling, SRPT was field-tested. In the next section, the demographic information of students, who participated in the field test, and the subsequent quantitative analysis of the revised SRPT items will be exhaustively discussed.

5.2 Quantitative analysis (2nd round) for field testing

5.2.1 Demographic characteristics of the field test participants

The field testing was conducted in six schools of Hadiya Zone, SNNPR. All the schools and teachers did not participate in the pilot test. The field testing was administered by the school teachers. The researcher had participated in facilitating the testing of grade 8 students. In the field test, 242 grade eight students from six schools had participated. Next, Rasch analysis that measures various psychometric properties of SRPT will be discussed.

5.2.2 Fit statics

In this section, fit statistics for the final version of SRPT will be discussed. A series of Rasch analysis procedures were presented to establish the psychometric properties of SRPT: fit statistics, unidimensionality, ICC, discrimination indices, Wright maps, and reliability computation. These analyses were used to evaluate whether SRPT items measure similar unidimensional construct – scientific reasoning, and eventually to identify the final items included in the SRPT. After evaluating the items in relation to the criteria of Rasch analysis, forty-four items (see Appendix A) were taken as the final items included measuring the construct scientific reasoning in physics for grade eight students.

The field testing of SRPT provided Infit MNSQ value of ZSTD for items 7.2, 7.3, and 7.4 were 4, 4, and 3.5 respectively, and outfit MNSQ value of ZSTD for items 7.2, 7.3, and 7.4 were 4, 3.2, and 3.3 respectively which indicate that the items could cause a threat for the validity of the test. The correlations of the items were also poor (less than 0.2) and which also indicates that the items are not in the same direction as the other items to contribute to measuring the construct scientific reasoning. The three items (item 7.2, 7.3, and 7.4) did not meet the criteria of Rasch analysis. Including the items in the test brought the tool inconsistent with the theoretically established levels of scientific reasoning. For this reason, item 7.1, 7.2, 7.3, and 7.4 were rejected from the final version of SRPT (item 7.1 was included in this list because the four items were developed in a way they are related to measuring the four levels of reasoning). Therefore, the analysis provided in the next sections will focus on the remaining forty items and the final version of SRPT comprises only forty items.

After rejecting the four items, the item fit metrics, which was reported in Infit MNSQ and Outfit MNSQ for the field test of 40 items, were .81 to 1.10 and .78 and 1.15 respectively which indicates that the learners' responses for the items are according to the levels established in the construct map. The ZSTD values for items for Infit and Outfit fit statistics for the field test were in the accepted interval (-2 to +2) except for items 3.2 and 4.3.

Items 4.3 and 3.2 were placed outside the expected range of ZSTD (see Figure 5.3). As the table in Appendix B has shown the fit statistics tell that item 4.3 has an infit MNSQ of 0.85, outfit MNSQ of .83, infit ZSTD of -3.3, and outfit ZSTD of -2.8. In the same way, item 3.2 has an infit MNSQ of 0.81, outfit MNSQ of .74, infit ZSTD of -3.6, and outfit ZSTD of -3.4. One can see from this that the items have acceptable MNSQ values and unacceptable ZSTD values. Such items are said to be overfitting items (Linacre, 2020) and don't cause any threat to the validity of the scale because they are measuring the same construct as far as the MNSQ value for infit and outfit lies within the expected range between 0.7–1.3 (Linacre, 2020; Martin & Jamieson-Proctor, 2019).

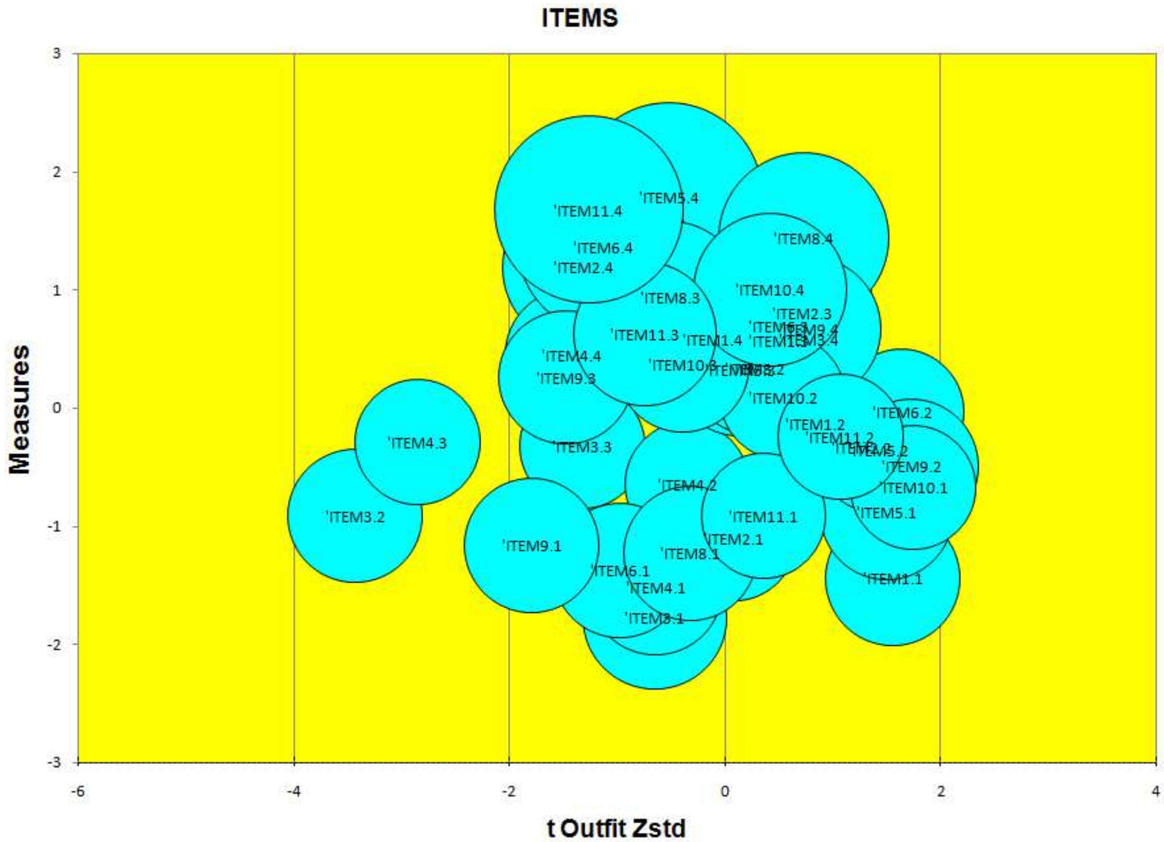


Figure 5.3 Bubble chart of the SRPT shows how each of the 40 items fit the scale

Furthermore, the validity of a scale becomes a threat when there are more misfit (underfitting) items on the right-hand side of the bubble chart than items (overfitting) on the left-hand side (Martin & Jamieson-Proctor, 2019). As shown in Figure 5.3, both items were placed on the left side of the bubble chart. Therefore, the two items don’t affect construct validity and remain as a final part of the SRPT.

Table 5.8: Infit and outfit values for the total students (242) and items (40)

		INFIT		OUTFIT	
	Mean	MNSQ	ZSTD	MNSQ	ZSTD
Items	17.8	1.00	.1	.98	-.1
Students	97.7	1.00	.0	.98	-.0

The Infit and outfit MNSQ overall values for 40 items were 1.00 and .98 respectively that indicates that the expected mean square value of 1.00 was achieved (Bond & Fox, 2007). The Infit and outfit ZSTD overall values for 40 items were .1 and -.1 respectively which also indicates that the expected ZSTD mean of 0 was almost achieved (see Table 5.8). According to the overall result of fit statistics, the levels of reasoning established in the construct map match students' responses for the items. Consequently, the results indicate that the data fit the model and this, in turn, implies that the overall fit of the SRPT reasonably represents the construct of scientific reasoning and all the items contribute to a single underlying construct (scientific reasoning).

5.2.3 Unidimensionality (PCA on residuals)

The assumption of unidimensionality is one of the most important aspects of Rasch analysis. Rasch Model assumes that the scale measures a single variable. The Unidimensionality of a test can be examined by observing patterns in the residuals. Residuals refer to the difference between the theoretical prediction and actual performance of the items to measure the expected trait (Bond & Fox, 2015). Data that contains smaller residuals are expected to be unidimensional. According to Reckase (1979), a measure is considered to be unidimensional when the Rasch model explains a minimum of 20% variance. For Embretson and Reise (2013) the criterion for unidimensionality is a minimum of 3:1 ratio of the variance explained by a Rasch measure to the variance of the first principal component of residuals. The unexplained variance in the first contrast is also expected not to be greater than 15%. Furthermore, Linacre (2020) has suggested that the first component of residual having an Eigenvalue less than 3 (strength of lesser than three items) probably unidimensional. With the help of Principal Component Analysis (PCA), the variance was examined to identify contrasts in the residuals. This enabled us to determine whether the data has a secondary dimension or not. Data with a secondary dimension contain items that measure constructs that were not planned to measure.

The unidimensionality of SRPT was evaluated based on four criterions:

- a measure is considered to be unidimensional when the Rasch model explains a minimum of 20% variance (Reckase, 1979);
- a minimum of 3:1 ratio of the variance explained by a Rasch measure to the variance of the first principal component of residuals is recommended (Embretson & Reise, 2013);
- The explained variance in 1st contrast is recommended not to be greater than 15% (Embretson & Reise, 2013);
- The unexplained variance of the Eigenvalue for the first contrast is recommended less than 3 (strength of lesser than three items) and unexplained variance explained by first contrast < 5% is acceptable (Linacre, 2020).

With the help of Principal Component Analysis (PCA), the variance was examined to identify contrasts in the residuals in the final version of SRPT.

Table 5.9: PCA Analysis of SRPT using the field test

Variance in Eigenvalue units	**Eigenvalue**	**Observed**	**Expected**
Raw variance explained by measures	12.4	23.7%	21.6%
Raw variance explained by person	3.6	6.9%	6.3%
Raw variance explained by items	8.8	16.8%	15.4%
Total raw variance in observations	52.4	100%	100%

Table 5.10 Contrast in Eigenvalue Units of SRPT

Contras in Eigenvalue units	**Eigenvalue**	**Observed**	**Expected**
Unexplained variance in 1st contrast	2.8	5.3%	7.0%

Unexplained variance in 2nd contrast	2.2	4.2%	5.5%
Unexplained variance in 3rd contrast	2.1	4.0%	5.3%
Unexplained variance in 4th contrast	1.9	3.6%	4.7%
Unexplained variance in 5th contrast	1.7	3.3%	4.3%
Raw unexplained variance (total)	40.0	76%	100%

In the final version of SRPT, the raw variance explained by the measures was 23.7% which fulfills the first criteria (see Table 5.9). Secondly, the ratio of the variance explained by measurement to the first contrast is about 5:1 which is also greater than the minimum criteria of 3:1. Thirdly, the unexplained variance in 1st contrast for SRPT was 5.3% which is also less than the recommended 15%. Finally, the unexplained variance accounted for Eigenvalues in the first 2.8, second, the third, the fourth, and the fifth contrast was 2.8, 2.2, 2.1, 1.9, and 1.7 respectively which were good according to the criteria (see Table 5.10).

This suggests that the forty items of SRPT are measuring the intended construct, scientific reasoning. This further shows that the levels of the construct map which was designed in a way to include the three types of knowledge in a progressive manner stand for measuring a single construct.

5.2.4 Item characteristics curve (ICC)

ICC is one parameter used to check the quality of the final version of SRPT. Theoretically, it is expected that the ICC curve and score of data nearly approximate and close to each other to show that the data fit the model. As it was displayed in Appendix G (see the ICC curves for the 40 items displayed in the Appendix G), most items fit the model. In this section, only two items (best fitting and least fitting) were presented for discussion purposes. In Figure 5.4 the red line is the item characteristic curve as expected by the Rasch model. It is the ideal score value. The Gray lines indicate the 95% confidence interval. The blue line represents the actual score value of the students for a given item. From Figure 5.4, one can see that the score of item 10.2 (blue line) is close to the ideal score, in the expected interval, and covered a meaningful distance of the ideal curve. This suggests that the item is a good fit for the model and is measuring what was supposed to measure according to the construct map. Figure 5.5 represents the total score on the test and the actual score of item 5.3. From Figure 5.5 one can see that the score of item 5.3 is far from the ideal score and there is a score outside the expected range, which means that the item is a less good fit for the model. This also suggests that the item was struggled to measure what was supposed to measure according to the construct map.

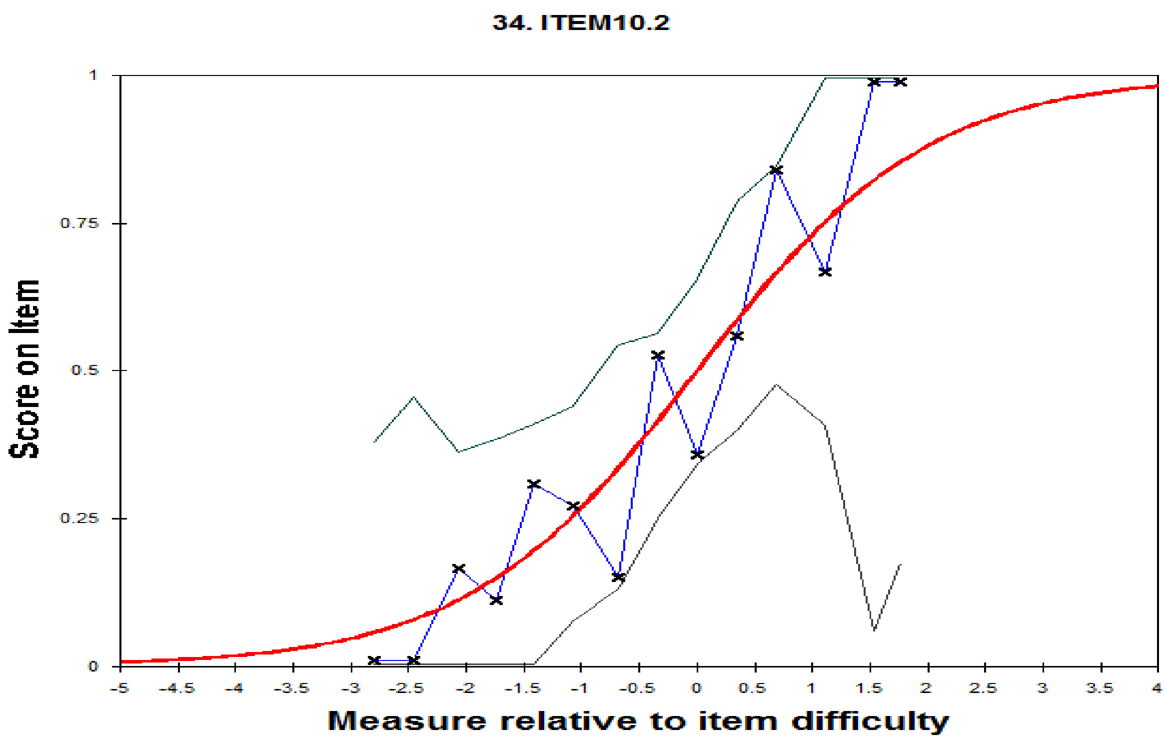


Figure 5.4: ICC curve for item 10.2 in terms of score and its difficulty

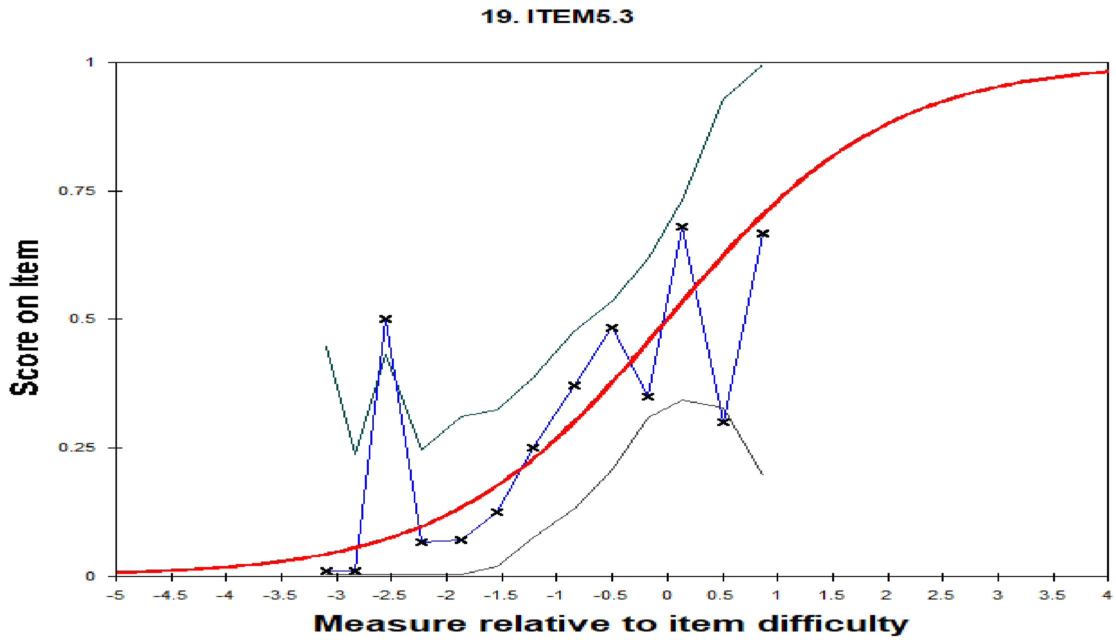


Figure 5.5: ICC curve for item 5.3 in terms of score and its difficulty

5.2.5 Discrimination: Category Values

The Rasch model is designed to provide the relation between the abilities of students and the difficulty of the items on the same scale. It can predict the responses of a person for a particular item by estimating the abilities level of the students in terms of the difficulty level of an item. It provides prediction whether a person can score right or wrong by estimating based on the ability of the person and the difficulty of the item. That is, it provides a way to identify whether more able persons have a greater likelihood of correctly answering all the items in a test and easier items are more likely to be answered correctly by all persons (Bond & Fox, 2015). More able students are expected to answer items correctly than less able students and easier items are expected to be answered by most students than difficult items. In the first version of SRPT, there were two items in which the score of less able students is greater than or equal to more able students. However, for the final version of SRPT, all the items have a greater probability to be answered by more able students than less able students. The average ability of students scoring 1 point (correct answer) for all the items is greater than students scoring 0 points (wrong answer) (see Appendix C). This shows that there is an improvement in terms of students' responses in the final version of SRPT and all the items have a positive correlation which tells that the items are pointing in the same direction to measure a single construct. This also indicates that the test is good enough to measure the construct of scientific reasoning appropriately as established in the construct map. The result further suggests that there is a linear relationship between students' overall score (achievement) and reasoning ability. That is, the most high able students achieve higher levels reasoning than the less able students. This strengthens the way scientific reasoning was conceptualized for this study in the way it consists of three forms of scientific knowledge progresses from content knowledge to epistemic knowledge.

5.2.6 Item person map

The Item-person map in Figure 5.6 represents item difficulty and person ability. To demonstrate such representation a logit scale is commonly used. The logit scale is a measurement unit to locate a person ability and item difficulty according to person's ability and item difficulty estimates (Bond & Fox, 2015). In Figure 5.6, the left side shows students and the right side shows items. Each "#" on the left side represents two students. The students at the top of the map are the high scoring and the items at the top of the map are the most difficult ones. A student 1 logit above or below an item indicates that the student has a 75% possibility to answer the item right or wrong respectively. A student placed opposite to the item on the map means the student has approximately 50% likelihood to answer the item correctly. The relative positions of item mean "M" and the person mean "M" provides the status of students' ability in terms of item difficulty.

The left side of Figure 5.6 shows the ability levels from the lower ability (*factual knowledge*) to the higher level (*drawing conclusion*). Students were distributed fairly from low abilities (at the bottom) to the low abilities (at the top). From Figure 5.6, one can see that only a few students' were able to achieve the highest level of scientific reasoning (*drawing conclusion*), and most students were able to achieve the lower levels (*factual knowledge and explanations of factual knowledge*). This is in agreement with qualitative findings (see Table 4.2, Table 4.4, Table 4.5, and Table 4.7).

From the right side of Figure 5.6, the item measure values on the item person map demonstrated that SRPT for all 40 items is ranged from the highest measure (item 5.4 with the measure of 1.79 logits) to the lowest measure (item 3.1 with the measure of −1.77 logits). From this figure, one can see that item 11.4 and 5.4 are the most difficult items. Item 11.4 is answered by only 28 (13%) students and item 5.4 is answered by 26 (11%) students correctly among 242 students. Item 8.4, item 6.4, item 2.4, item 10.4, item 2.3, and item 8.3 are also very difficult items for most students. They are good items to measure high able students but still, they are very demanding. Item 3.1 is the easiest item for the students and answered by 178 (75%) students correctly. Other items such as item 6.1, item 4.1, item 1.11, item 9.1, item 8.1, item 3.2, item 11.1, and item 4.2 are ordered in the bottom part of the item person map which is according to the theoretical prediction except for item 3.2 and item 4.2. In the same way, item 11.4, item 5.4, item 8.4, item 6.4, item 2.4, item 10.4, item 2.3, and item 8.3 are also located in the upper part of the item person map except item 8.3 and 2.3 which is also according to theoretical prediction in Table 2.5. Items 11.3, 6.3, 8.2, 1.4, 1.3, 3.4, 4.4, 9.4, 9.3, and 5.3 are items for measuring evidence level of reasoning. This is also according to theoretical predications except some items were included from the level of the drawing the conclusion. Finally, item 10.2, item 10.3, item 6.2, item 9.3, item 1.2, item 5.2, item 11.2, item 2.2, item 3.3, item 9.2, item 10.1, item 4.2, item 4.3, and item 10.1 are located around the level of explanation which is also according to expected prediction except for a few items. This indicates that the items which measure a higher level of reasoning such as the evidence generation and the drawing conclusion were answered by a few students and items which require low ability such as recalling facts were answered by most students. Students' ability and item difficulty were distributed in terms of what has been expected according to the levels generated in Table 2.5. Note that the items ending with .1 represent level 1, .2 represent level 2, .3 represent level 3, and .4 represent level 4.

The item-person map also demonstrated the items and students are fairly distributed in the scale except there were gaps among items between item 3.1 and item 4.1; item 8.4 and item 11.4. The location of item 3.1 was -1.77 (ability measure) and the location of item 4.1 was -1.52. The location of item 8.4 was 1.44 and the location of item 11.4 was 1.68. The gaps between the items were less than 0.3 logits. However, the distance between the items in logit value implies that there is no significant gap among the items that force to modification of the items. This suggests that students' ability was well targeted with the scale.

The item measures for SRPT were expressed using the logit scale that ranged from −1.77 for item 3.1 to 1.79 for item 5.4. The person measures for SRPT were expressed using the logit scale that ranged from −2.77 to 1.3. The range of the person's ability is wider at the bottom of the item person map and thinner in the upper part of the map. The range of person measure in logit value is greater than the item measure. This indicates, along with the position of students mean (M) and the item mean (M), that the test was difficult for the sample and it is required more easy items to represent the item difficulty and person ability in better proportion.

The result from the item person map, generally, indicates that the scientific reasoning test seems to fairly measure students' reasoning ability as predicted in the construct map. The item person map also indicates that the items and persons are consistently distributed. This suggests that SRPT provides evidence for construct validity and that the instrument is measuring in a way that matches what the theory predicted.

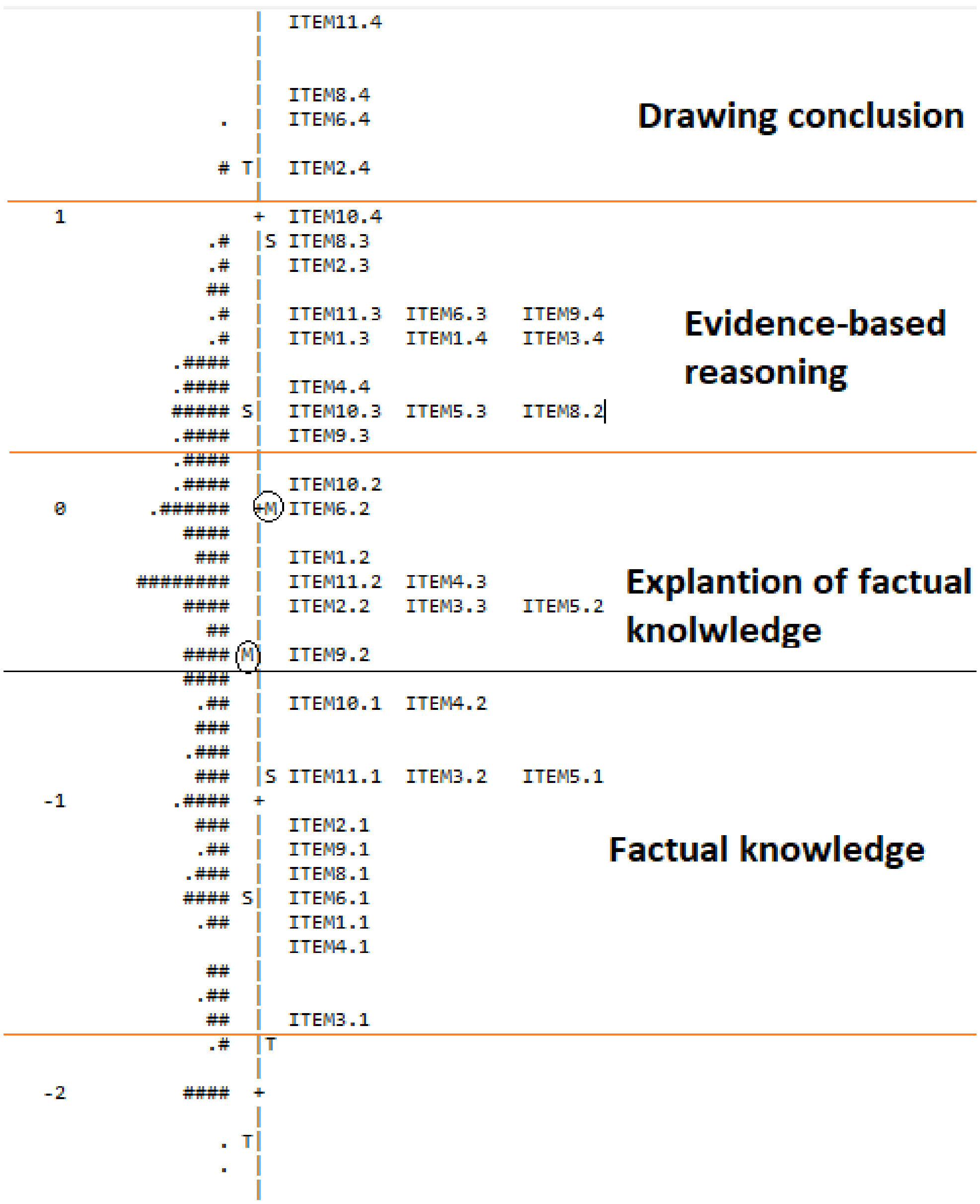


Figure 5.6 Item person map for the SRPT: Logit scale with increasing values from bottom to top distributes the person abilities (on the left) and the item difficulties (on the right).

5.2.7 Reliability Calculation

The separation and reliability index tells the reproducibility of the items and orderings of persons. The item separation is used to determine how well the items are ordered to separate the students in their ability levels. The item separation (> 3) implies the item hierarchy is good enough to verify the item difficulty into high, medium, and low; item reliability (> .9) implies the person sample is large enough to distinguish the high, medium, and low achievers (Linacre, 2020).

Table 5.11: Summary of Person and item Statistics: Reliability Coefficients and Separation Indices

				INFIT		OUTFIT	
	Mean	Separation	Reliability	MNSQ	ZSTD	MNSQ	ZSTD
Items	15.6	5.75	.97	1.00	.0	.98	.0
Students	94.3	1.88	.78	1.00	.0	.98	-.1

The summary statistics of 40 items and 242 students are shown in Table 5.11. The result of item separation of SRPT (5.75) along with the item reliability (.97) confirms that the person sample is large enough to separate item difficulty hierarchy to measure the expected levels of reasoning. The item reliability (.97) along with item separation (5.75) indicates that the items were hierarchically separating the ability of students into four levels of reasoning (Boone et al., 2013). The item reliability (.97) suggests that the SRPT is replicable. That means if one gives the test to another similar group; it can be replicated on this order of item estimation.

The person reliability (.78) along with person separation (1.88) indicates that the SRPT tool separates students into high, medium, and low achievers (Boone et al., 2013). The person reliability (.78) also suggests that when one gives these persons other similar tests the item can estimate reliably. The data, in general, suggests that the SRPT has the quality to separate students into groups based on their ability, and the items are ordered in the desired continuum based on the difficulty levels.

The final version of SRPT was comprised of 40 items developed from the grade eight physics syllabus. The developed assessment tool, SRPT, differs from international tools such as TIMSS, PISA, and national assessments such as NLA, regional assessments, and classroom assessments in many regards. Some of the unique features of SRPT includes the following:

a) SRPT is context-based. In developing and validating the SRPT items, qualitative data was collected from students and their teachers using interviews and FGD respectively. In addition, students reasoning pattern was explored and teachers were engaged in the developments, revision and validation stage of SRPT. Hence, the interview of grade eight students and the iterative discussion and engagement of teachers shaped SRPT to be contextual;

b) The construct being studied, scientific reasoning, was conceptualized using construct map before item creation, in a manner that incorporate the three types of knowledge (content, procedural and epistemic) across students' learning progression from low level of reasoning to high level of reasoning (factual knowledge, explanation of factual knowledge, evidence-based reasoning, and drawing conclusion);

c) The tool development and validation followed an iterative process which enabled to modify the construct map and items;

d) Most of the available test instruments that attempted to measure scientific reasoning did not implemented the Rasch model in developing and validating their tests. However, SRPT, in addition to qualitative validation, followed construct-driven assessment procedure by using Rasch model for the analysis and validation stages.

To sum up the chapter, it was found from the quantitative result that SRPT is valid and reliable to measure students' scientific reasoning ability. After making some modifications on the items, item and person separation were improved (see Table 5.11), and items were ordered according to the construct map (see Figure 5.6). This implies that the revised version of SRPT (see Appendix A) is more valid and reliable to measures the construct scientific reasoning than the first version (pilot testing). Based on the findings of chapter 4 and chapter 5, the next chapter will present discussions of the results of this study.

Chapter 6 Discussions, conclusions and recommendations

6.1 Discussions

This dissertation aimed to develop and validate a scientific reasoning assessment tool that was used to investigate students' patterns and progress of scientific reasoning for grade eight students in physics in the Ethiopian context. The following research questions guided the study:

1) Which scientific reasoning skills are addressed by the construct map developed for the SRPT test items?
2) How are the content and construct validity of the Scientific Reasoning Progress Tool (SRPT) items qualitatively established?
3) How effective are the SRPT test items in measuring grade eight students' scientific reasoning in physics?
4) What is the trend of Grade 8 students' progression among the different reasoning subskills measured by the SRPT?

In the previous chapters (chapters 4 and 5), data were analyzed qualitatively and quantitatively. In this chapter, the result of the analysis will be discussed in terms of qualitative and quantitative findings vis-à-vis of literature to answer the research questions. The trend of Grade 8 students' progression among the different scientific reasoning sub-skills will be discussed in terms of naive, factual knowledge, explanation of factual knowledge, evidence-based reasoning, and drawing conclusions. Based on the findings of this study and literature, the research questions will specifically be answered and finally, conclusions and recommendations will be followed.

6.1.1 Discussions of general findings

6.1.1.1 Naïve reasoning

From chapter 4, it was found that students come into the class with various non-scientific conceptions. According to findings from phenomenography, all the students were observed that they hold a naive view about physics concepts in one form or another. There was a dominant belief that science presented in the textbooks, reference books, and classroom teachings were considered as knowledge to be taken as it is. Some students believed that physics is difficult to understand, universal truth, and unchangeable (see section 4.1). It was also revealed that most students come into class with the assumption that concepts of physics are not related with day to day experiences. These might force students to think of science or physics in the beginning based on the conception they have from the environment around them. Such reasoning most of the time is not scientific. Such perceptions may lead students to undermine advanced studies, thinking, and reasoning and could have an effect on students' low conceptual understanding and reasoning abilities.

Students' naïve reasoning is well-established type of knowledge in science education. There is theoretical and empirical evidence that suggests students come into classes with various inconsistent conceptions with scientific knowledge. Theoretically, it was confirmed that students come into class with various inconsistent conceptions about science (Croker & Buchanan, 2011; Schauble, 1996). Some considered this naïve reasoning as a common-sense belief about physics, based on daily experiences (Bertamini, Spooner, & Hecht, 2004) and it is described as the way of thinking shared commonly by most people instead of thinking scientifically. Others considered these as irrelevant, incomplete, and meaningless responses (Ford & Wargo, 2012). Some others are also have found that students believed physics is difficult to understand, abstract, and imaginary (Verspoor, 2008). Such students' view was found to be one factor that can affect students' scientific knowledge construction and higher-level reasoning (Croker & Buchanan, 2011; Verspoor, 2008).

One of the most challenges related to students' scientific knowledge construction and high-level reasoning, according to Hammer and Elby (2003), is considering students' naïve conceptions of physics as fixed, stable, and context general. They suggested that students' naïve conceptions of physics should be viewed as productive epistemic resources that can be used to assist students build new knowledge and higher-level reasoning through effective instruction. This implies that it is critical to respect what students have in their minds before introducing formal instruction, rather than pressuring them to discard their thinking system. This allows students to learn scientific knowledge and develop higher-level reasoning without having to strive or exert unnecessary effort. This liberates the minds of anyone concerned about students' knowledge construction by demonstrating that there is no need to destroy information in order to replace it with new knowledge; rather, there is a need to go from one point of view to another based on acceptable evidence. Students learn physics more effectively when instructions are based on context-sensitive resources and students understand physics knowledge as a system of ideas that can be developed gradually by re-constructing and refining current understanding rather than thinking of physics knowledge as a coherent system that can be acquired from external sources such as as teachers, texts and scientists according to reasearch studies (Hammer & Elby, 2003).

6.1.1.2 Factual knowledge

From the students' interview, FGD with primary school physics teachers, and quantitative Rasch analysis it was found that students do not faced difficulty in attaining factual knowledge. About 91% of students attained this level (see Table 6.1). It was easy for most of the students to state laws and theories, identify units of measurement and formulae, read data from the tables and graphs, and differentiate variables (see Table 6.1).

Table 6.1 Summary of students' achievement per the levels reasoning for SRPT

Levels of reasoning	Number of students achieved the level	Percentile of the students achieved the levels (%)
Factual knowledge	222	91
Explanation of factual level	137	56
Evidence-based reasoning	56	23
Drawing conclusion	3	1.2

However, as it was found from students' interviews and FGD the focus of middle school science is on factual learning. Verspoor (2008) and Joshi and Verspoor (2013), strengthens this finding which contends that the classroom assessments and instructions of sub-Saharan African countries including Ethiopia are dominated by students' low-level abilities. Such a trend may limit students' further development in thinking and reasoning abilities. This, in turn, may lead students to develop a culture of status quo instead of developing practical problem-solving abilities for complex situations.

Table 6.2 demonstrates the key findings for the factual level along with the indicative responses for the categories. It was found that students believed facts, concepts, and formulae in textbooks are the basic foundation for advanced knowledge and important to construct new knowledge. This is in agreement with the literature. This level in literature is considered a low-level understanding of science. Because it is the cornerstone for the building of all scientific knowledge, there are no reasons for skipping this level. Before establishing conceptual knowledge, students must first master the fundamentals. In literature, this level is considered as an inevitable part of school science for students' healthy and proper progress towards higher levels like evidence generation and evaluation (Anderson, 2012; Ford & Wargo, 2012). This suggests that students can't provide proper explanations, engage in constructive thinking and arguments, and make a progress into high-level reasoning without having a proper foundation based upon the basics of factual knowledge.

Table 6.2: Categories vs indicative responses of factual level

Level of reasoning:	**Factual**	
Key categories	Indicative responses	Descriptions
Scientific fact	...I first try to remember definitions and formulae to understand better	Facts, concepts, and formulae in the text books are important to construct new knowledge.
Mixed fact	...facts are observations and concepts discovered by scientists. ...is reliable knowledge... (student 12)	Physics facts are the product of men's concepts and observations about physical phenomenon and they are also reliable
Naïve fact	$g = 9.8\ m/s^2$ is true value of gravity.	Concepts, laws and theories are discovered by physicists to be recalled and remembered.

From the students' interviews, it was also explored that they believed that facts, concepts, and formulae are truth but not to be subjected to change and need to be taken as they are written in the text. When students are confronted with scientific knowledge at the surface level they start to define, recall, and remember physics concepts. At this stage, students faced the difficulty of accepting anything outside textbook definitions, data recorded in the tables, and values in the graphs. Facts are considered to be basics to establish science knowledge and are all about learning physics that is similar to Schommer's (1990) findings in which students considered knowledge of facts, formulae, and procedural solving strategies as a comprehensive understanding of physics knowledge. There is also students' tendency to capture more concepts by rote learning than trying to construct a deeper conceptual understanding about a specific concept and this could be related to curriculum influence that enforces students to cover a wide range of contents. This was also reflected in FGD with primary school teachers. It was found that students, soon the lesson starts, prioritize copying notes from the blackboard than participating in a teacher's discussion and engaging themselves in group activities; they attempt to answer questions from the textbooks and repeat what teachers said as it is. This view about facts limits students' scientific knowledge construction and developing reasoning at a higher level (Schommer, 1990).

It was also found that physics facts are considered by students as concepts, laws, and theories discovered by physicists to be recalled and remembered. Once knowledge is discovered by someone and it works forever. Such belief tends to a simplistic view of science. This was also consistent with research findings in which students think that scientific knowledge is objective and universal (Hansson & Leden, 2016). This belief is related to students' views about the nature of science. For such students, science and the law of nature are the same. This suggests that it is required to consider the nature of science in physics education and curriculum materials to help students be informed about the nature of scientific knowledge (Hansson & Leden, 2016; Lederman, 2006; Matthews, 2012). Students with such view of science knowledge compel to develop naive reasoning (Bertamini et al., 2004).

6.1.1.3 Explanation of factual knowledge

The findings revealed that grade eight students' scientific reasoning ability is limited majorly to the first and second levels of reasoning (see Table 6.1). This was reflected in quantitative results in which most students were heavily populated in the two lower levels of reasoning (about 222 (91%) attained level one and 137 (56%) achieved level two). Item person map also demonstrated that more than two-thirds of students achieved the first and second level of scientific reasoning. i.e. most students answered only the items that were developed for first and second levels.

After having some more understanding of physics concepts, students start to explain and reflect their understanding based on the knowledge they got. Before students conceptualize the concept, they faced the difficulty of creating an integration between what they have in their minds with what they are introduced in the physics class. This process of creating a link between their prior knowledge and newly introduced physics knowledge was observed as a challenging moment for students. At this level, students were observed actively trying to give meanings by providing related examples. This stage is characterized by students tending to use the knowledge they grasped at the factual level to relate concepts with daily practices, provide explanations, and integrate concepts to give meanings for a physical phenomenon. It was observed that students tend to relate factual knowledge with the natural phenomenon to construct science knowledge. Literature also supports this level characterizing by students' attempt to relate particular facts and concepts to arrive at the general conception of how those facts and concepts give meanings (Corcoran et al., 2009). Ford and Wargo (2012) put this level as a level of explanation and characterized by students recount scientific knowledge to explain the natural phenomena. This view about explanations of scientific knowledge is expected to lead students in constructing higher-level knowledge and reasoning (Corcoran et al., 2009).

Table 6.3: Categories vs indicative responses of level of explanation

Level of reasoning: Explanation		
Key categories	Indicative responses	**Description**
Scientific explanation	Velocity......is the distance covered from starting to end point in a specific direction.... (explanation) (student 12)	An explanation is provided by relating different variable
Mixed explanation	... velocity represents how a fast a body moves *(*explanation (student 6)	There is difficulty of providing scientific explanation by relating different variables.
Naïve explanation	F = ma is law of nature discovered by scientists (explanation) (student 5)	Physics explanation is absolute.

Table 6.3 demonstrates the key findings for the explanation of factual knowledge along with the indicative responses for the categories. It was found that students attempt to construct science knowledge by relating with day to day experiences without catching scientifically accepted knowledge. They tend to make use of science concepts to establish prior belief without having an understanding of the very meaning of the concept. This is consistent with research findings in that students try to give meaning to physics

concepts unscientifically based on the experiences they faced regularly (Esanu & Hatu, 2015). The reason could be students tend to memorize facts, formulae, and problem-solving procedures instead of trying to construct deep understanding (Elby, 1999) that stems from naïve beliefs about the nature of physics, and persistence to maintain unscientific beliefs (Chinn & Malhotra, 2002) and lack of depth understanding of the concept (Harlen, 2013). They do not seem ready to accept scientific conceptions against prior knowledge in the beginning. students fail to establish the scientific conceptualizations because they attempt to construct knowledge by focusing on only pursuing good grades that is also in agreement with the literature (Elby, 1999; Verspoor, 2008). The results also indicate that students are struggling to transcribe what they learned into materializations to create meaningful scientific conceptualizations. The result implies that the majority of grade eight students have a limited understanding of how scientific knowledge is constructed which is also consistent with literature that reveals that most k-8 students' have faced such difficulty (Carey & Smith, 1993). This could be one of the reasons that students face the difficulty of generating evidence-based reasoning (Corcoran et al., 2009).

There was also a tendency to seek a flawless explanation. They think that providing an explanation is discovering naturally aligned working principles for various things in common. This thinking could stem from the students' views of the nature of science. This is also consistent with the literature in which students try to construct universal scientific knowledge by assuming there is only a single method and way of constructing scientific knowledge and which is more procedural than creative (Hansson & Leden, 2016). They don't understand that scientific knowledge can be constructed using various ways and it is tentative in nature.

6.1.1.4 Evidence-based reasoning

Evidence-generation is the most critical stage for students' real progress towards higher reasoning ability. At the level of explanation, students conceptualize what seems meaningful for them until they are confronted with evidence. It was found that a few students recognized that one needs to check with proper evidence whether a claim works or not before accepting or rejecting (see Table 6.1). This is in agreement with literature which suggests that evidence in science is used either to support or reject claims (Kind, 2013; McNeill & Krajcik, 2011; Sampson & Clark, 2008). Once students provide a claim, generating evidence to support the claim is what follows and establishment of scientific reasoning (Berland & McNeill, 2010).

Table 6.4: Categories vs indicative responses of evidence generation

Level of reasoning: Evidence generation		
Key categories	Indicative responses	**Descriptions**
Scientific evidence	The evidence for the presence of gravity is falling objects (student 2)	Evidences can be obtained from physics concepts, laws and theories
Mixed evidence	The sources of evidences are teachers, experienced persons and natural phenomenon...	Unable to identify valid and invalid evidences
Naïve evidence	... ...physics laws and theories are true evidence (evidence) (student 9)	Physics concepts, laws, theories are true sources of evidences.

As it was found from students' interviews that one of the challenges with generating evidence was the tendency to use mixed evidence (partially correct). Students were generating evidence from day to day experiences to accept or reject someone's claim. They were trying to generate evidence for explanations from examples they observe in the environment around them but the evidence was not scientifically accepted (See Table 6.4). However, one can help students generate evidence, accept or reject a claim by providing physics concepts in a way that students can touch, feel, and demonstrate, and practice based on their environment (Erlina, Susantini, Wasis, & Pandiangan, 2018) which was found to be one of the missing elements of the schools' science.

It was also explored from students' interview that they believed that evidence is required to support a claim but the significance of evidence is to catch an infallible truth which has already been proved by scientists. For most students, the evidence is only to understand and apply on the daily basis the knowledge which has already been established in physics theories, laws, and principles once and for all. Most students failed to understand that evidence is either to accept or reject a claim tentatively until the best alternatives views emerge.

Previous studies revealed that type of instruction and assessment strategies are the main ones in hindering students' development of evidence-based reasoning. Most instruction are not designed in a way to promote higher learning abilities and incorporate a simplistic form of inquiry; assessments are also exam-oriented and meant to measure merely factual understanding (Anderman et al., 2012). Another challenge for students' lack of ability to generate scientific evidence could be related to the approach of classroom teaching which focuses on helping students to pass the regional exams than providing activities that invite students to be engaged in evidence-based activities that are in agreement with previous studies (Anderman et al., 2012; Chinn & Malhotra, 2002; Joshi & Verspoor, 2013). The type of assessment provided by teachers also contributes to students' poor ability in generating evidence. The assessments developed by teacher's most of the time meant to measure factual; they lack assessing students' ability to generate evidence which also agrees with previous findings (Anderman et al., 2012; Joshi & Verspoor, 2013).

It was also found that students tend to provide evidence to support the mental conceptualization but the quality and quantity of evidence they provide is not convincing enough. As it was revealed from data, students prefer solving problems procedurally than engaging themselves in evidence-based discussions and activities which might be the reason why the students fail to generate valid evidence that is in agreement with Schommer (1990).

Quantitative results indicated that students tend to generate evidence if the concepts are provided in clear, context-based, and authentic activities. It was observed from both qualitative and quantitative study students faced difficulty in generating valid evidence. The type of evidence generated mostly unrelated to scientific knowledge, mixed with day to activities, and sometimes irrelevant. About 23% of the students tried to generate valid evidence

6.1.1.5 Drawing conclusions

This level was observed that the most difficult level for the students to achieve (See Table 6.1). The qualitative and quantitative study revealed that students have faced difficulty to draw a conclusion based on valid scientific evidence. They faced a challenge in reflecting the nature of the tentativeness of scientific conclusions and unable to draw scientific conclusions according to the results from the SRPT analysis. Only 1.2% (see Table 6.1) of the students were able to attain this level. They are more tempted to draw a conclusion based on the intuitive knowledge and notion of the universal nature of scientific knowledge. The conclusions are mostly invalid and stem from the evidence generated from day to day experiences which are not consistent with the recent scientific research findings. Rasch item-person map indicates that only a few students' abilities were matched with the items developed to measure students' ability of drawing conclusion. Mostly, students tried to generate evidence to establish their pre-existing knowledge instead of attempting to draw new conclusions. This implies that students' way of constructing knowledge, nature of evidence, and conclusions are heavily influenced by naïve reasoning.

Table 6.5 depicts various students' views about drawing conclusions. There are a few students' who believed the evidence-based conclusion and they were attempting to draw a conclusion based on the evidence from scientific principles, laws, and theories. They believed that one can arrive at a currently acceptable conclusion based on evidence, discussions, and arguments. They understood that based on the available evidence scientific knowledge can be modified, rejected, and accepted. Literature suggested that this is a level with the requirement of students' ability to connect the conceptualizations (claims and explanations) with evidence to arrive at a conclusion (Kind, 2013). Providing a claim and supporting the claim with the evidence is not always final, uncertainty and disagreement about their relationship may occur and need to be justified to draw conclusions (Kind, 2013; Sampson & Clark, 2008).

Table 6.5: Categories vs indicative responses of drawing conclusion

Level of reasoning Drawing conclusion		
Key categories	Indicative responses	
Scientific conclusion	....F = ma can be improved (student 1)	Scientific conclusions are tentative and subject to change and modifications
Mixed conclusion	....based on Newton's first law of motion plants are at rest and will remain at rest for ever (student 5)	Lack of ability to state a law physics (partial understanding of law of physics)
Naïve conclusion	......findings are truth discovered by scientists (student 11)	One can arrive on an absolute conclusion

For some students, a conclusion is to arrive at an ever-working principle of nature for a specific phenomenon (See Table 6.5). They concluded that, if someone arrives at a conclusion about a phenomenon based on valid evidence, there would be no change in such findings. Such beliefs are related to the conception of the existence of universal and constant knowledge which is the product of the unchangeable nature of physics concepts, theories, and laws (Hansson & Leden, 2016).

It was difficult for students to draw conclusions scientifically. Why students fail to draw valid conclusion could be related with such skills requires high cognitive abilities and practices in a regular basis (Erlina et al., 2018) and requires training and teacher's skills (Han, 2013) which were not the trend of the schools as it was obtained in this study. Children face difficulty in generating evidence, drawing conclusions, and other types of reasoning skills unless the concepts are contextualized, necessary scaffoldings are provided and activities are devised properly that can help them encourage to participate (Alfieri, Brooks, Aldrich, & Tenenbaum, 2011; Butler & Markman, 2012; Zimmerman, 2007).

Some literature on students' understandings of the nature of science (NOS) characterizes three levels of epistemic development (Carey & Smith, 1993; Ford & Wargo, 2012; Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002). At the lowest level, students can carry out scientific reasoning by remembering, describing, and solving conceptual problems. At this level students' reasoning ability is characterized by holding the positivist view that distinguishes little between describing and explaining nature. At an intermediate level, students can explain their reasoning and provide evidence for the explanations. It requires knowledge about science and methods. At this level, students tend to express the possibility of multiple explanations for science phenomena, but they may acquire a view that there exists a single reality. At the most advanced level, students can critique scientific reasoning and draw a scientific conclusion based on the available evidence. According to the result of the current study, the students' reasoning ability is limited towards the lowest level, and intermediate level to a lesser extent. They faced difficulty in attaining higher epistemic levels.

Studies have also revealed that a lack of high-level reasoning is related to students' positivist view which propagates objective reality (Niaz, 2017). The dominance of school science with objectivist teaching strategies which is based on the positivist epistemology, a dominant teaching method in sub-Saharan African countries (Teshome, 2017; Verspoor, 2008), influences school science to be dependent on remembering what is written, factual memories, and algorithmic learning. This might imply that students think that learning physics is recalling and remembering facts, repeating what was said instead of constructing knowledge, and developing reasoning based on the newly emerging data resources. Research findings have also shown that students with a naïve view of science focus on factual knowledge and accept scientific knowledge without reasoning, and questioning, accept everything written in the textbook and what is told by teachers, and they are unable to relate scientific knowledge with day to day practices (Edmondson & Novak, 1993; Özdemir, 2007).

This section provided general discussion about the result of this study. In the following sections, the answers for the research questions will specifically be presented.

6.1.2 Research question one

1) Which scientific reasoning skills are addressed by the construct map developed for the SRPT test items?

Research question one explores the nature and characteristics of students' scientific reasoning and elucidates the construct map that guides the development of SRPT items. This research question further investigated students' views across the four levels of scientific reasoning, which were identified through the reviews of relevant literature..

In this dissertation, interviews with 12 grade eight students from both urban and rural districts around Hosanna town were used to understand the nature of students' scientific reasoning. The on-site interview with students had lasted from 10 to 20 minutes. The interview was guided by six semi-structured interview questions (given in Appendix E).

It was found from the students' interviews that the students had demonstrated three different views, namely *naïve view, mixed view, and scientific view* for each of the four-level of scientific reasoning, namely *factual knowledge, explanation of factual knowledge, evidence-based reasoning, and drawing conclusion*. The results from the qualitative analysis of students' interviews revealed that the majority of students had shown naïve views across the four levels of reasoning. Moreover, as the complexity of the levels of scientific reasoning increases (say from *factual knowledge to drawing conclusion),* the naivety of students' views also increases. Most of the students had shown naïve views (75 %) for the fourth level of scientific reasoning (drawing conclusion), while 33 % of students were naïve for the first level of scientific reasoning (factual knowledge). All in all, naïve view is the dominant view of students' understanding of the nature of science.

The development of the construct map was bounded around students' interviews, teachers' FGD, and relevant literature. The interview and FGD data were transcribed, read thoroughly using relevant documents again and again, and collected memos and explanations. On top of this, Ph.D. candidates in Physics Education from Addis Ababa University were recruited in auditing the data analysis to ensure the credibility of the result.

Moreover, iteratively consulting the findings of students' interviews, the results of teachers FGD, the collected relevant literature, and the conceptual frameworks depicted in Table 2.5 had yielded the person-item map of grade eight students' scientific reasoning in physics. The person-item map provided a comprehensive account of the level of grade eight students' scientific reasoning skills along with the students' views about each level of scientific reasoning. The findings also imply that the students' level of reasoning ranges from scientific view to mixed view and then to naïve view, whereas the complex level of scientific reasoning range increase from factual knowledge to drawing conclusion. Students based on their reasoning ability can be categorized into three strata: Level 1 (naïve), Level 2 (mixed), and Level 3 (scientific), whereas the scientific reasoning construct has four hierarchical that spread in the continuum from factual knowledge (Level 1), explanation of factual knowledge (Level 2), evidence-based reasoning (Level 3), and drawing conclusion (Level 4).

The developed person-item map in terms of the students' view is shown in Figure 4.1. This person-item map guided the development of the initial items. In the SRPT items development phase, the due emphasis was given to these matrices to include items that assess content, procedural, and epistemic knowledge of students across the corresponding proficiency levels along with the students' view towards the four levels of scientific reasoning. The next research question answers how the content and construct validity were qualitatively established.

6.1.3 Research question two

2) How are the content and construct validity of the Scientific Reasoning Progress Tool (SRPT) items qualitatively established?

Teachers' FGD and review of SRPT items by the panel of experts were used to ensure validity qualitatively. In both, teachers' FGD and experts reviews, items that have vague wordings, lack of contextualization into the students' ability level, problems related to the scientific reasoning contents, language problems, and lack of clarity were identified and thoroughly revised based on the suggestions provided to improve the problematic SRPT items. During the revision, 31 items had been revised, and 5 items were replaced by new items. Some exemplars had been provided under Section 4.2 and 4.3. The validations of teachers and experts were critical in establishing the content as well as construct validity of SRPT, which in turn provide evidence to claim that SRPT is a valid assessment tool. Moreover, the qualitative validation phase ensured content validity by qualitatively evaluating how the SRPT items are valid to measure the concept of scientific reasoning (the construct). The iterative reviews and revisions ensured that the qualitative methods of validation had established content validity as well as construct validity. In addition to the qualitative methods utilized, this dissertation had also explored the quantitative psychometric analysis of SRPT items using Rasch modelling. In the next section, this study used quantitative analysis to answer research questions three and four.

6.1.4 Research question three

3) How effective are the SRPT test items in measuring grade eight students' scientific reasoning in physics?

One of the main tasks of this study was developing and validating the SRPT. The tool was, initially developed and content validity was assured qualitatively. These test items were administered for the students and the construct validity was assured quantitatively. It was found that the SRPT is offered as a valid and reliable instrument that is a construct-driven, domain-specific, and higher-order thinking assessment that can measure students' reasoning progress. To ensure the construct validity of the SRPT, Rasch modelling was used. In terms of Rasch indicators such as fit statistics all the forty items of SRPT fell in the acceptable range (Table 5.8), and satisfies the minimum criteria of unidimensionality (see the description in Section 5.2.3), items were distributed on the item person map according to the predicted order (see Figure 5.6). Furthermore, item and person reliability, and item and person separation were found to be in the acceptable range (see Table 5.11). This indicates that the SRPT classified students' from low-level reasoning ability (factual knowledge) to high-level reasoning ability (drawing conclusion); it separated students into low, medium, and high achievers; it was valid and reliable. This further indicates that the SRPT was appropriate and fairly measures students' scientific reasoning progress.

6.1.5 Research question Four

4) What is the trend of Grade 8 students' progression among the different reasoning subskills measured by the SRPT?

This study dealt with students' reasoning abilities in relation to the four levels scientific reasoning progress. The students' trend of progression in relation to the four reasoning level of scientific reasoning presented as follows:

Factual knowledge should not be neglected in science learning since it is the foundation for the construction of other levels of reasoning progresses. Students are required to grasp the basics before establishing conceptual understanding. In literature, this level is considered as an inevitable part of school science for students' healthy and proper progress towards higher levels like evidence generation and evaluation (Anderson, 2012; Ford & Wargo, 2012). The items meant to measure students' lower reasoning abilities such as remembering factual knowledge, identifying units of measurement, identifying the given values, identifying unknown values were answered by most students and ordered in the item-person map according to the theoretical construct map. The students' responses were in agreement with Ford and Wargo (2012) in which they categorized this as a low level of students' conceptual understanding. At this level students are expected to put things as they are; that is why the students did not struggle to answer the items and the items are properly measuring the students' abilities.

The level of explanation is characterized by students' attempting to relate particular facts and concepts in order to arrive at a general conception of how those facts and concepts give meanings (Corcoran et al., 2009). It was found that the items which were designed to measure the level of explanation are also ordered as expected and established in the level of reasoning progress in

Table 4.10. At this stage, the items were developed to measure students' abilities to relate concepts or variables to provide meaningful explanations. The tool seems to properly measure students' ability. Most students were able to provide correct explanations for the factual responses. Students were also able to solve physics problems using formulae, and provided meanings for the data, and tried to relate physics theories with day to day activities to some extent.

Evidence Generation is the most important stage in students' actual progress toward improved reasoning ability. Even if the tool works satisfactorily, students were observed struggling to generate evidence. Even if generating evidence is the core component of scientific reasoning (Kind, 2013; McNeill & Krajcik, 2012; Sampson & Clark, 2008), it was demanding for most grade 8 students' to support the claim with valid scientific evidence. Only a few students were able to generate scientific evidence for the explanation or claims. The reason why students faced the difficulty of generating valid evidence might be related to which evidences require one's creative abilities to match the theoretical relationship between concepts, laws, and theories with available data (Kuhn, 2002; Zimmerman, 2000). To provide evidence, students have also required the knowledge of practical tasks such as experimental activities, day to day activities, and demonstrations (Zimmerman, 2000). Another challenge for students' lack of ability to generate scientific evidence could be related to the approach of classroom teaching which focuses on helping students to pass the regional exams more than providing activities that invite students to be engaged in evidence-based activities, which is in agreement with previous studies (Anderman et al., 2012; Chinn & Malhotra, 2002; Joshi & Verspoor, 2013). The type of assessment provided by teachers also contributes to students' poor ability in generating evidence. The assessments developed by teachers mostly are meant to measure factual and conceptual knowledge; it lacks assessing students' ability to generate evidence, which also in line with previous findings (Anderman et al., 2012; Teshome, 2017).

The items that were meant to measure students' ability to *draw conclusions* were also ordered at the top of the item-person map indicating that most students were not able to draw a scientific conclusion. It was found that only a few students were able to draw conclusions scientifically. The reason for students' low ability to draw conclusions could be related to a lack of conceptual understanding and the inability to generate valid evidence. In order to draw a conclusion, students need the ability to compare and contrast evidence from various sources. Why students fail to draw valid conclusions could also be due to lack of practices of such skills in regular classroom activities, and skills of drawing conclusions require high cognitive abilities (Erlina et al., 2018), and requires continuous training and teachers' skills (Han, 2013). Children face difficulty in generating evidence and drawing conclusions unless the concepts are contextualized, necessary scaffoldings are provided and activities are devised properly to encourage them in participatory tasks (Alfieri et al., 2011; Butler & Markman, 2012; Zimmerman, 2007).

Finally, this study has produced two important contributions to science education. Firstly, a valid and reliable assessment tool (SRPT) was developed that could help to measure students' scientific reasoning ability (see appendix A). Secondly, a theoretical model was developed (see Table 5.7) that could depict students' ability of scientific reasoning, guide assessment development in scientific reasoning, and syllabus and curricular materials developers.

6.2 Conclusions

This study was engaged in exploring students' reasoning patterns, and the development and validation of scientific reasoning progress tool for grade 8 students based on Senocak's and Wilson's assessment framework. It revealed that students hold a positivist view about scientific reasoning progress levels. Students think that scientific knowledge is static, fixed, universal, certain, and unchangeable.

The study obtained that SRPT is valid and reliable; it classified students' from low-level reasoning ability to high-level reasoning ability; it separated students into low, medium, and a high achievers. The items are ordered according to theoretical prediction – easiest items ordered in the lower part of the construct map and the most difficult items are ordered in the upper part; most students were able to answer the easiest items and very few students were able to answer the most difficult items – which is expected according to the theoretical model.

It was also found that grade 8 students' scientific reasoning ability was poor which was demonstrated from the item person map in which the mean value of the items is greater than the mean value of the student. Most students were able to achieve only the first levels; they faced difficulty in attaining higher reasoning levels.

Studies revealed that lack of high-level reasoning and students' naive is related to students' positivist view which propagates objective reality (Niaz, 2017; Özdemir, 2007). The dominance of school science with objectivist teaching strategies which is based on the positivist epistemology, a dominant teaching method in Sub-Saharan African countries (Teshome, 2017; Verspoor, 2008) influences school science to be dependent on remembering what is written, factual memories, and algorithmic learning. This might imply that students think that learning physics is recalling and remembering facts, repeating what was said instead of constructing knowledge, and developing reasoning based on the newly emerging data resources. Research findings have also shown that students with naïve view of science focus on factual knowledge, they accept scientific knowledge without reasoning and questioning, accept everything written in the textbook and what is told by teachers, and they are unable to relate scientific knowledge with day to day practices (Edmondson & Novak, 1993; Özdemir, 2007).

6.3 Recommendations

Based on the findings of this study and research findings the following recommendations were forwarded for curriculum designers, classroom teachers, test developers, and science education researchers.

6.3.1 The need for higher level understanding knowledge and reasoning assessments

Students' reasoning ability is expected to progress when they are engaged in a complex cognitive task such as evidence-based argumentation (Kuhn, 2005), promoting inquisitive learning strategy (Harlen, 2013), employing higher ability assessment (Opitz et al., 2017), and employing evidence-driven tools (Corcoran et al., 2009). Ethiopian government launched an education and training policy which demands focus on problem-solving and student-centered teaching to improve the quality of education and tackle the old-age teacher-centered approach by implementing the various program (MoE, 1994). However, all the efforts made were not by considering evidence-based students learning progression and evidence-based tools that tell how students' conceptual understanding and reasoning progresses varies. The objectives are meant to cover a wide range of contents, memorizing facts, solve procedural problems, demanding responses without detailed explanations and evidence, and arrive at non-evidence-based conclusions. From this study, it was also found that there is a students' tendency to memorize more concepts than trying to construct a deeper conceptual understanding and engaging themselves in evidence-based activities. This could be one of the indicators of curriculum influences (Joshi & Verspoor, 2013) which compels students to learn more concepts than in-depth learning. Even if the Ethiopian curriculum framework encourages higher-order learning and assessment, the classroom assessments are dominated by lower-order knowledge assessments. Therefore, there is a need to consider inquiry-based teaching not only in curriculum materials but also in classroom situations. There is a need to devise a mechanism to follow up the classroom situations about students learning towards inquisitive learning strategy.

Studies revealed that one of the challenges that retard the development of students' reasoning abilities is the nature of instruction and activities provided for the students. The over-simplistic and a simple way of answering and questioning method dominated in middle school science. Chinn and Malhotra (2002) questioned overly simplistic forms of inquiry as an obstacle to promote higher reasoning abilities. They further suggested that classroom instructions that do not incorporate high-level reasoning that requires to think about alternative explanations or to connect theory to data lead students to algorithmic reasoning. The assessments of the schools are exam-oriented either to pass or fail students and to measure merely factual memory. The recommended remedy to overcome such problems is introducing dialogical teaching in a way that allows students to talk to each other meaningfully, provide evidence and justify their evidence along with proper assessment mechanisms (Adey & Csapo, 2012). Therefore, it is required to consider and shift the classroom assessment system, national assessments, and textbook assessments towards evidence-based and higher-order ability measuring such as evidence generation and drawing conclusions. To enhance students' reasoning skills it is required to shift the current assessment system which heavily focuses on the content-based ability by incorporating assessments that measure higher reasoning abilities in the instructions. The curriculum needs to emphasize inquiry-based science education (Dunbar & Klahr, 2012; Han, 2013), dialogical teaching (Osborne, 2010; Osborne, Erduran, & Simon, 2004) along with the proper assessment strategies to enhance students' higher reasoning abilities.

6.3.2 The need for incorporating three forms of scientific knowledge

The teaching goals in STEM education also need to include enhancing the three types of knowledge: content, procedural, and epistemic knowledge. This study indicated that students' are limited in the factual and to some extent in the explanation level. This suggests that there is a need to develop content, procedural, and epistemic knowledge to act in higher-order thinking which is also supported by research findings (Kind, 2013; Kind & Osborne, 2017). Merely focusing on content-based teaching and assessments inhibit students' long-term academic achievement in science and personal development. To shift the current dominance of science education from emphasizing content-oriented teaching methods and lower order cognitive demands which produces students with mostly factual knowledge to construct-based and higher-order cognitive demands there is a need to shift classroom instruction along with the assessment system (Kambeyo, 2017; Osborne, 2013).

6.3.3 The need for inquiry based instruction in consideration with the NOS

Even if the students' reasoning ability was limited majorly in the lower levels, there is an indication from the result which suggests that students' reasoning ability can be improved towards higher-order reasoning if suitable instructions and assessment techniques are preferred that can promote such abilities. The curriculum needs to emphasize inquiry-based science education with explicit instruction of the nature of science (NOS) as an instructional approach (Dunbar & Klahr, 2012; Duschl & Grandy, 2011; Meyer & Crawford, 2011), dialogical teaching (Osborne, 2010; Osborne et al., 2004) along with the proper assessment strategies to enhance students' higher reasoning abilities. It is recommendable that teacher training institutions and curriculum designers need to incorporate the nature of science in the school curriculum and classroom instructions explicitly to enhance teachers' and students' views about the nature of science. Students should be informed of the modern view of the nature of science which emphasizes that scientific knowledge is tentative; empirical, derived from human inference, imagination, and creativity; socially and culturally embedded (Lederman, 2006; NRC, 2001). Explicit instruction of the nature of science in classroom instruction enables students to think that scientific knowledge is subjective, tentative, socially constructed, culturally embedded instead of thinking scientific knowledge as objective reality, factual information, universal truth, and to be transferred from one source to another (Meyer & Crawford, 2011). To enhance students' reasoning skills it is also required to shift the current assessment system which heavily focuses on content-based and low ability by incorporating instructional activities and assessments which measures higher reasoning abilities during classroom instructions.

This study revealed that most students do not understand science or physics as a way of knowing. They only understand physics as something which exists naturally and waits for somebody to reveal to them. Students perceived physics as merely concept, law, and theory instead of thinking and realizing physics as a culture of practicing and exercising activity to construct tentative knowledge. Such a view may lead to physics as a skill to accomplish a task not an enterprise to construct new knowledge. The problem stems from students' views about the nature of science. This problem could be overcome by incorporating explicitly the nature of science and the history of physics in science education and school science. Introducing the history of physics – particularly incorporating a short introduction about how a particular theory or law originated and take current form can help to inform students to appreciate that physics is simply a progressive activity practiced over the years by individual scientists.

6.3.4 The need to contextualize science concepts

The other finding of this study was students have faced difficulty in relating scientific concept with day to day activities to develop meaningful understanding and scientific reasoning. One of the important aspects of science education needs to be making science relevant, practical, and meaningful for students. It is important to help students to create a proper link between day-to-day activities and scientific concepts. This can be achieved by providing proper examples from daily experiences and everyday phenomenon that students observe and experience regularly (Ng & Nguyen, 2006). Educators and concerning bodies need to think about how to apply scientific knowledge to real life situations. This needs to be reflected in science education and curriculum materials to inform experts and teachers to incorporate in classroom instruction.

This study provides evidence-based students' reasoning progress, students' pattern of responses, and level of understanding and reasoning. This will provides insights for curriculum developers on what needs to be incorporated in the curricular materials, how objectives are re-arranged, and what objectives to be incorporated. It also provides some inputs for a classroom teacher what instruction plan is appropriate, what types of activities are required, what types of assessment is appropriate, what kind of scaffolding is required, what instructional materials are suitable, how students' reasoning ability to be developed and validated, and with what prior conceptions and pattern of responses students might come into the classroom.

Finally, this study followed a construct-driven assessment model to assess and develop scientific reasoning abilities and tools. It was observed in this study that the Rasch model is a simple and effective tool for the analysis of the quality of items and in the development of scales. The design of this research may be adapted to develop additional instruments to investigate the learning progression of students' scientific reasoning in middle and high school classes in Ethiopian and other African countries.

6.4 Limitations of the study

This study focuses only on the development and validation of scientific reasoning progress tools that can measure in a limited aspect of students' ability of scientific reasoning. The study provides evidence which implies the middle school students' reasoning progresses from the generation of claim to the drawing conclusion. However, a single study in limited schools and a single construct can't provide the whole picture of students' status and levels of reasoning. More studies are required to explore students' pattern of reasoning progresses to obtain detailed information about students reasoning abilities and identify possible naïve reasoning which enables to make appropriate improvements in the current learning methods and assessment strategies.

This study was conducted to explore students' patterns of reasoning based on the four levels identified in the literature. However, it was found that there is a need for more levels to represent and assess students' reasoning ability in a better way. The result identified a gap in the lower level to represent students' abilities. Including more levels (especially in the lower levels) along with suitable items can provide a way to represent more students and their performance.

The study did not incorporate classroom intervention. It is highly required to conduct classroom intervention to identify various alternative conceptions, students' reasoning progresses, and patterns of reasoning based on suitable instruction to help students' knowledge construction and develop higher reasoning abilities.

References

Adey, P., & Csapo, B. (2012). Developing and assessing scientific reasoning. In B. Csapo & G. Szabo (Eds.), *Framework for diagnostic assessment of science* (pp. 17-49). Budapest: Hungary: NemzetiTankonyvkiado.

Al Fallay, I. S. (2018). Test Specifications and Blueprints: Reality and Expectations. *International Journal of Instruction, 11*(1), 195-210. doi: https://doi.org/10.12973/iji.2018.11114a

Alemu, M., Kind, P., Tadesse, M., Atnafu, M., & Michael, K. (2017). Challenges of Science Teacher Education in Low-income Nations - The Case of Ethiopia *PART 13: STRAND 13*, 1782.

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? . *Journal of Educational Psychology, 103*(1), 1-18. doi:https://doi.org/10.1037/a0021017

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education, 93*(3), 389-421.

Altinok, N., Angrist, N., & Patrinos, H. A. (2018). *Global dataset on education quality (1965-2015)*. Washington: DC World Bank.

Anderman, E. M., Sinatra, G. M., & Gray, D. L. (2012). The challenges of teaching and learning about science in the twenty-first century: Exploring the abilities and constraints of adolescent learners. *Studies in Science Education, 48*(1), 89-117.

Andersen, C., & Garcia-Mila, M. (2017). Scientific reasoning during inquiry. In K. S. Taber & B. Alpan (Eds.), *Science education. new directions in mathematics and science education* (pp. 105–117). Rotterdam: The Netherlands: Sense.

Anderson, K. J. (2012). Science education and test-based accountability: Reviewing their relationship and exploring implications for future policy. *Science Education, 96*(1), 104-129.

Arega, M. (2014). *Classroom Assessment Manual for Primary and Secondary School Teachers.* . Addis Ababa, Ethiopia: NEAEA.

Au, W. (2007). High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis. *Educational researcher, 36*(5), 258-267. Retrieved from https://journals.sagepub.com/doi/abs/10.3102/0013189X07306523. doi:10.3102/0013189x07306523

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transaction, 22*(1), 1145-1146.

Bao, L., Cai, T., Koenig, K., Fang, K., Han, J., Wang, J., . . . Luo, Y. (2009). Learning and scientific reasoning. *Science*, *323*(5914), 586-587.

Bell, S. (1997). Learning with information systems: Learning cycles in information systems development. In F. Marton & S. Booth (Eds.), *Learning and Awareness*. Mahwah, NJ: Lawrence Earlbaum.

Berland, L., & McNeill, K. (2010). A learning progression for scientific argumentation: Understanding student work and designing supportive instructional contexts. *Science Education, 94*(5), 765-793.

Bertamini, M., Spooner, A., & Hecht, H. (2004). The representation of naïve knowledge about physics. In G. Malcolm (Ed.), *Multidisciplinary Approaches to Visual Representations and Interpretations* (pp. 27-36): Elsevier.

Bilican, K., Cakiroglu, J., & Oztekin, C. (2015). How contextualized learning settings enhance meaningful nature of science understanding. *Science Education International*, *26*(4), 463-487.

Billingsley, B., & Fraser, S. (2018). Towards an understanding of epistemic insight: The nature of science in real world contexts and a multidisciplinary arena. *Research in Science Education, 48*(6), 1107–1113.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139 - 148.

Bless, C., Higson-Smith, C., & Kagee, A. (2006). *Fundamentals of social research methods. An African perspective* (4th ed.). Cape Town, South Africa: Juta.

Bloem, S. (2013). *PISA in Low and Middle Income Countries OECD Education Working Papers, No. 93*. Paris, France: OECD Publishing Retrieved from http://dx.doi.org/10.1787/5k41tm2gx2vd-en.

Bond, T., & Fox, C. (2007). *Fundamental measurement in the human sciences*: Institute for Objective Measurement.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model : Fundamental measurement in the human sciences* (3rd ed.). New York, NY: Routledge.

Boone, W., Staver, J., & Yale, M. (2013). *Rasch analysis in the human sciences*: Springer.

Brown, N. J. S., & Wilson, M. (2011). A Model of Cognition: The Missing Cornerstone of Assessment. *Educational Psychology Review, 23*(2), 221. Retrieved from https://doi.org/10.1007/s10648-011-9161-z. doi:10.1007/s10648-011-9161-z

Burton, L. J., & Mazerolle, S. M. (2011). Survey instrument validity part I: Principles of survey instrument development and validation in athletic training education research. *Athletic Training Education Journal, 6*(1), 27-35.

Butler, L. P., & Markman, E. M. (2012). Finding the cause: Verbal framing helps children extract causal evidence embedded in a complex scene. *Journal of cognition and development, 13*(1), 38-66.

Carey, S., & Smith, C. (1993). On understanding the nature of scientific knowledge. *Educational Psychologist*, *28*(3), 235-251.

Cavanagh, R., & Sparrow, L. (2010a). *Measuring mathematics anxiety: Paper 2-Constructing and validating the measure.* Paper presented at the AARE 2010 Conference Proceedings.

Cavanagh, R., & Sparrow, L. (2010b). *Measuring Mathematics Anxiety: Paper 2 – Constructing and Validating the Measure*. Paper presented at the Australian Association for Research in Education Conference, Melbourne, Australia.

Champagne, A. (2015). Assessment: An overview. In R. Gunstone (Ed.), *Encyclopedia ofScience Education* (pp. 87-89). New York, NY: Springer Science+Business Media

Chinn, C., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education, 86*(2), 175-218.

Chmielewski, A. K., & Reardon, S. F. (2016). Patterns of Cross-National Variation in the Association Between Income and Academic Achievement. *AERA Open*, *2*(3), 1-27.

Constantina, S., & Constantine, S. (2014). Subjectivity and objectivity in science: An educational approach. Advances in Historical Studies, 3, 183-193. In.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform*. Philadelphi, PA:: Consortium for Policy Research in Education.

Cresswell, J., Schwantner, U., & Waters, C. (2015). A review of international large-scale assessments in education: Assessing component skills and collecting contextual data.

Creswell, J. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Boston: Pearson.

Creswell, J., & Clark, V. (2018). *Designing and conducting mixed methods research* (3rd ed.). Los Angeles, CA: Sage.

Creswell, J., & Guetterman, T. (2019). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (6th ed.). New York, NY: Pearson.

Creswell, J., & Poth, C. (2018). *Qualitative Inquiry & Research Design Choosing Among Five Approaches* (4th ed.). Los Angeles, CA: Sage Publications.

Croker, S., & Buchanan, H. (2011). Scientific reasoning in a real-world context: The effect of prior belief and outcome on children's hypothesis-testing strategies. *British Journal of Developmental Psychology, 29*(3), 409-424.

Crombie, A. C. (1994). *Styles of scientific thinking in the European tradition: The history of argument and explanation especially in the mathematical and biomedical sciences and arts*. London, UK: Duckworth.

Davies, M. (2013). Critical thinking and the disciplines reconsidered. *Higher Education Research & Development, 32*(4), 529-544.

DeBoer, G. E. (2011). The globalization of science education. *Journal of Research in Science Teaching, 48*(6), 567-591.

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. Retrieved from https://doi.org/10.1007/s10763-017-9844-0

Dole, S., Bloom, L., & Kowalske, K. (2016). Transforming pedagogy: Changing perspectives from teacher-centered to learner-centered. *Interdisciplinary Journal of Problem-Based Learning, 10*(1), 1.

Dornyei, Z. (2007). *Research methods in applied linguistics*. Cambridge, UK: Oxford University Press.

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education, 84*, 287 – 312.

Dunbar, K. N., & Klahr, D. (2012). Scientific thinking and reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 701-718). Oxford, UK: Oxford University Press.

Duschl, R. A., & Grandy, R. (2011). Demarcation in science education: Toward an enhanced view of scientific method. In R. S. Taylor & M. Ferrari (Eds.), *Epistemology and Science Education: understanding the evolution vs. intelligent design controversy* (pp. 3-19). New York, NY: Routledge.

Edmondson, K. M., & Novak, J. D. (1993). The interplay of scientific epistemological views, learning strategies, and attitudes of college students. *Journal of Research in Science Teaching, 30*(6), 547-559.

Elby, A. (1999). Another reason that physics students learn by rote. *American Journal of Physics, 67*(S1), S52-S57.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*: Psychology Press.

Engelhard Jr, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*: Routledge.

Erlina, N., Susantini, E., Wasis, W., & Pandiangan, P. (2018). The Effectiveness Of Evidence-Based Reasoning In Inquiry-Based Physics Teaching To Increase Students' Scientific Reasoning. *Journal of Baltic Science Education, 17*(6), 972.

Esanu, A., & Hatu, C. (2015). *The significance of prior knowledge in physics learning.* Paper presented at the The International Scientific Conference eLearning and Software for Education.

Fernald, L. C., Kariger, P., Engle, P., & Raikes, A. (2009). *Examining early child development in low-income countries: a toolkit for the assessment of children in the first five years of life*: World Bank.

Fernandes, G. W. R., Rodrigues, A. M., & Ferreira, C. A. (2018). Conceptions of the nature of science and technology: A study with children and youths in a non-formal science and technology education setting. *Research in Science Education, 48*(5), 1071–1106.

Fischer, F., Chinn, C. A., Engelmann, K., & Osborne, J. (2018). *Scientific reasoning and argumentation: The roles of domain-specific and domain-general knowledge*: Routledge.

Fischer, F., Kollar, I., S., U., Sodian, B., Hussmann, H., Pekrun, R., . . . Eberle, J. (2014). Scientific reasoning and argumentation: advancing an interdisciplinary research agenda in education. Frontline Learning Research, 2(3), 28–45. *Frontline Learning Research*, *2*(3), 28–45.

Ford, M. (2015). Educational implications of choosing “practice” to describe science in the Next Generation Science Standards. *Science Education*, *99*(6), 1041–1048.

Ford, M., & Wargo, B. M. (2012). Dialogic framing of scientific content for conceptual and epistemic understanding. *Science Education, 96*(3), 369-391.

Ford, M. J., & Forman, E. A. (2006). Chapter 1: Redefining disciplinary learning in classroom contexts. *Review of Research in Education, 30*(1), 1-32.

Gelman, S. A. (2003). *Oxford series in cognitive development. The essential child: Origins of essentialism in everyday thought*: Oxford University Press.

Gelman, S. A., & Noles, N. S. (2011). Domains and naïve theories. *Wiley Interdisciplinary Reviews: Cognitive Science, 2*(5), 490-502.

Glewwe, P., & Kremer, M. (2006). Chapter 16 Schools, Teachers, and Education Outcomes in Developing Countries. In E. Hanushek & F. Welch (Eds.), *Handbook of the Economics of Education* (Vol. 2, pp. 945-1017): Elsevier.

Goldman, S. R., & Pellegrino, J. W. (2015). Research on Learning and Instruction. Implications for Curriculum, Instruction, and Assessment. *Policy Insights from the Behavioral and Brain Sciences, 2*(1), 33-41. doi: 10.1177/2372732215601866

Gott, R., Duggan, S., & Roberts, R. (2008). Concepts of evidence. *School of education: University of Durham*.

Greaney, V., & Kellaghan, T. (2008). *Assessing national achievement levels in education, National assessments of educational achievement Volume 1.* . Washington DC: The World Bank.

Griffin, S., & Case, R. (1997). Wrap-Up: Using peer commentaries to enhance models of mathematics teaching and learning. *Issues in Education, 3*(1), 115-134.

Hacking, I. (1992). 'Style' for historians and philosophers. *Studies in History and Philosophy of Science Part A, 23*(1), 1-20. Retrieved from http://www.sciencedirect.com/science/article/pii/003936819290024Z. doi:https://doi.org/10.1016/0039-3681(92)90024-Z

Hacking, I. (2002). Historical ontology. In *In the Scope of Logic, Methodology and Philosophy of Science* (pp. 583-600): Springer.

Hacking, I. (2012). 'Language, truth and reason'30 years later. *Studies in History and Philosophy of Science Part A, 43*(4), 599-609.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education, 15*(3), 309-333.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38-47. doi:10.1111/j.1745-3992.1993.tb00543.x

Hammer, D., & Elby, A. (2003). Tapping epistemological resources for learning physics. *Journal of the Learning Sciences, 12*(1), 53-90.

Han, J. (2013). *Scientific reasoning, Research, Development and Assessment*. Colombus, Ohio: Ohio State University Press.

Hansson, L., & Leden, L. (2016). Working with the nature of science in physics class: turning 'ordinary' classroom situations into nature of science learning situations. *Physics Education, 51*(5), 055001. Retrieved from http://dx.doi.org/10.1088/0031-9120/51/5/055001. doi:10.1088/0031-9120/51/5/055001

Hanushek, E. A., & Woessmann, L. (2015). *The Knowledge Capital of Nations: Education and the Economics of Growth*. Cambridge, MA: MIT Press.

Harlen, W. (2013). *Assessment & inquiry-based science education: Issues in policy and practice*: Global Network of Science Academies.

Hewitt, P. (2006). *Conceptual Physics* (10 ed.). San Franscisco, CA: Pearson: Addison Wesley.

Hill, C. (2008). The post-scientific society. *Issues in Science and Technology on Line, 24*(1), 78 – 84.

Joshi, R., & Verspoor, A. (2013). *Secondary Education in Ethiopia: Supporting Growth and Transformation*. Washington, DC: World Bank.

Kambeyo, L. (2017). Scientific Reasoning Skills: A Theoretical Background On Science Education. *NERA Journal, 14*, 40-64.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of educational Measurement, 38*(4), 319-342.

Kane, M. T. (2006). *Validation. In R. L. Brennan (Ed.), Educational measurement*. New York, NY: American Council on Education/Macmillan.

Kennedy, C., & Wilson, M. (2007). Using progress variables to interpret student achievement and progress (BEAR Report Series, 2006-12-01). *University of California, Berkeley*.

Keys, C. W. (1999). Revitalizing instruction in scientific genres: Connecting knowledge production with writing to learn in science. *Science Education, 83*(2), 115-130.

Kind, P. (2013). Establishing Assessment Scales Using a Novel Disciplinary Rationale for Scientific Reasoning. *Journal of Research in Science Teaching, 50*(5), 530-560.

Kind, P., & Osborne, J. (2017). Styles of Scientific Reasoning: A Cultural Rationale for Science Education? *Science Education, 101*(1), 8–31.

Kivunja, C., & Kuyini, A. B. (2017). Understanding and applying research paradigms in educational contexts. *International Journal of higher education, 6*(5), 26-41.

Kuhn, D. (2002). What is scientific thinking, and how does it develop? In U. Goswami (Ed.), *Handbook of childhood cognitive development* (pp. 371-393). Oxford, UK: Blackwell.

Kuhn, D. (2005). *Education for thinking*: Harvard University Press.

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

Kuhn, D., Schauble, L., & Garcia-Mila, M. (1992). Cross-Domain Development of Scientific Reasoning. *Cognition and Instruction, 9*(4), 285-327. doi:https://doi.org/10.1207/s1532690xci0904_1

Lamb, S., Jackson, J., & Rumberger, R. (2015). ISCY technical paper: Measuring 21st century skills in ISCY.

Lawson, A. E. (2004). The nature and development of scientific reasoning: A synthetic view I. *International Journal of Science and Mathematics Education, 2*(4), 307–338. Retrieved from https://link.springer.com/article/10.1007%2Fs10763-004-3224-2.

Lederman, N. (2006). Syntax of nature of science within inquiry and science instruction. In *Scientific inquiry and nature of science* (pp. 301-317): Springer.

Lederman, N., Abd-El-Khalick, F., Bell, R. L., & Schwartz, R. S. (2002). Views of nature of science questionnaire: Toward valid and meaningful assessment of learners' conceptions of nature of science. *Journal of Research in Science Teaching, 39*(6), 497-521.

Lehrer, R., & Schauble, L. (2015). The development of scientific thinking. In L. Liben, U. Muller, & R. Lerner (Eds.), *Handbook of Child Psychology and Developmental Science* (Vol. 7th, pp. 671- 714). Hoboken, NJ: Wiley.

Linacre, J. M. (2020). A User's Guide to Winsteps and Ministep Rasch-Model Computer Programs. Chicago: Beaverton, Oregon. Retrieved from https://www.winsteps.com/manuals.htm

Linn, R. L., & Baker, E. L. (1995). What do international assessments imply for world-class standards? *Educational Evaluation and Policy Analysis, 17*(4), 405-418.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ Lawrence Erlbaum Associates.

Martin, D., & Jamieson-Proctor, R. (2019). Development and validation of a survey instrument for measuring pre-service teachers’ pedagogical content knowledge. *International Journal of Research & Method in Education*, 1-14.

Marton, F. (1986). Phenomenography: A research approach to investigating different understandings of reality. *Journal of Thought, 21*(3), 28-49.

Marton, F., & Booth, S. (1997). *Learning and Awareness*. Lawrence Erlbaum.

Martone, A., & Sireci, S. G. (2009). Evaluating Alignment Between Curriculum, Assessment, and Instruction. *AERA, 79*(4), 1332–1361. Retrieved from http://rer.aera.net.

Matthews, M. R. (2012). Changing the focus: from nature of science (NOS) to features of science (FOS) In M. S. Khine (Ed.), *Advances in Nature of Science Research: Concepts and Methodologies* (pp. 3–26). Dordrecht: Springer.

McNeill, K., & Krajcik, J. (2011). Supporting Grade 5-8 Students in Constructing Explanations in Science: The Claim, Evidence, and Reasoning Framework for Talk and Writing. *Pearson.*

McNeill, K., & Krajcik, J. (2012). Book study facilitator's guide: Supporting grade 5–8 students in constructing explanations in science: The claim, evidence and reasoning framework for talk and writing. In: New York: Pearson Allyn & Bacon.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators.Research, 45*, 35–44.

Meyer, J., & Land, R. (2003). *Threshold concepts and troublesome knowledge: Linkages to ways of thinking and practicing within the disciplines (ETL Occasional Report 4).* Edinburgh, Scotland: School of Education, University of Edinburgh.

Meyer, X., & Crawford, B. A. (2011). Teaching science as a cultural way of knowing: Merging authentic inquiry, nature of science, and multicultural strategies. *Cultural Studies of Science Education, 6*(3), 525-547.

Mislevy, R. J. (2007). Validity by design. *Educational Researcher, 36*, 463–469. doi:10.3102/0013189X07311660

MoE. (1994). *Education and Training Policy*. Addis Ababa, Ethiopia: St. George Printing Press.

MoE. (2003). *Teacher Education System Overhaul [TESO] Handbook.* Addis Ababa, Ethiopia: Ministry of Education.

MoE. (2005). *Educational Sector Development Program III (ESDP III)*. Addis Ababa, Ethiopia: Ministry of Education.

MoE. (2009). *Curriculum Framework for Ethiopian Education (KG - Grade 12)*. Addis Ababa, Ethiopia: Ministry of Education.

MoE. (2019). *Education Analysis of National Learning Assessment Systems Ethiopia Country Report*. Addis Ababa, Ethiopia: Ministry of Education.

MoE & JICA. (2016). *Strategic Policy for National Science, Technology and Mathematics Education*. Addis Ababa, Ethiopia: MoE & JICA Retrieved from http://www.moe.gov.et/documents/20182/42694/Strategic+Policy+for+National+Science+Technology+and+Mathematics+Education_Full.pdf/ef372bc1-8bc3-430a-b561-b6535e95419a?version=1.0.

Moore, J. C., & Rubbo, L. J. (2011). Scientific reasoning abilities of non-science majors in physics-basedcourses. Retrieved from arXiv:1110.2764v1 [physics.ed-ph].

Mullis, I. V., Martin, M. O., & Sainsbury, M. (2016). PIRLS 2016 reading framework. *PIRLS*, 11-29.

Nagaoka, J., Farrington, C. A., Ehrlich, S. B., & Heath, R. D. (2015). *Foundations for Young Adult Success: A Developmental Framework. Concept Paper for Research and Practice*: ERIC.

NAGB. (2015). *Science Framework for the 2015 National Assessment of Educational Progress: Policy Statement*. Washington, DC.

Ng, W., & Nguyen, T. V. (2006). Investigating the integration of everyday phenomena and practical work in physics teaching in Vietnamese high schools. *International Education Journal, 7*(1), 36-50.

Niaz, M. (2017). *Evolving Nature of Objectivity in the History of Science and Its Implications for Science Education*. Cham, Switzerland: Springer.

NRC. (2000). *How people learn: Brain, mind, experience, and school: Expanded edition*: National Academies Press.

NRC. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

NRC. (2007). *Taking science to school: Learning and teaching science in grades K-8.* Washington, DC: National Academies Press.

NRC. (2011). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. Committee on a Conceptual Framework for New K-12 Science Education Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education*. Washington, DC: The National Academies Press.

NRC. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.

Nunan, D. (1992). *Research Methods in Language Learning*. Cambridge, UK Cambridge University Press.

OECD. (2011). *Education at a Glance 2011: OECD Indicators*. Paris, France: OECD Publishing Retrieved from http://dx.doi.org/10.1787/eag-2011-en.

OECD. (2012). *Education at a Glance 2012: Highlights*. Paris, Frances: OECD Publishing.

OECD. (2013). *Draft PISA 2015 Science Framework.* Paris, France: OECD Publishing Retrieved from http://www/oecd.org/pisa/pisaproducts/pisa2015draftframeworks.htm.

OECD. (2015). *Education at a Glance 2015: OECD Indicators*. Paris, France: OECD Publishing Retrieved from http://dx.doi.org/10.1787/eag-2015-en.

OECD. (2016). *Education at a Glance 2016: OECD Indicators*. Paris, France: OECD Publishing Retrieved from http://dx.doi.org/10.187/eag-2016-en.

OECD. (2017). *PISA for Development Assessment and Analytical Framework: Reading, Mathematics andScience, Preliminary Version*. Paris, France: OECD Publishing.

Oh, H., Seo, D., & Kozub, F. (2010). The Emotional Reactions to Challenging Behavior Scale-Korean (ERCBS-K): Modification and Validation. *Adapted Physical Activity Quarterly, 27* (1), 17–31.

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

Oriondo, L. L., & Dallo-Antonio. (1998). *Evaluating educational outcomes (test, measurement, and evaluation)* (5th ed.). Quezon City: REX Printing Company, Inc. .

Osborne, J. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *Science, 328*(5977), 463 – 466.

Osborne, J. (2013). The 21st century challenge for science education: Assessing scientific reasoning. *Thinking Skills and Creativity, 10*, 265-279.

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching, 41*(10), 994 - 1020.

Osborne, J., Rafanelli, S., & Kind, P. (2018). Toward a more coherent model for science education than the crosscutting concepts of the next generation science standards: The affordances of styles of reasoning. *Journal of Research in Science Teaching, 55*(7), 1-20.

Osborne, R. J., & Wittrock, M. C. (1985). The generative learning model and its implications for science education. *Studies in Science Education, 12*, 59 - 87.

Özdemir, G. (2007). The effects of the nature of science beliefs on science teaching and learning. *Uludağ Üniversitesi Eğitim Fakültesi Dergisi, 20*(2), 355-372.

Pollitt, E. (2001). The developmental and probabilistic nature of the functional consequences of iron-deficiency anemia in children. *J. Nutr., 131*(2), 669S-675S.

Prince, M. (2004). Does Active Learning Work? A review of the Research. *Journal of Engineering Education, 93*(3), 223-231.

Rasch, G. (1960). *Probabilistic model for some intelligence and achievement tests.* . Copenhagen, Denmark: Danish Institute for Educational Research.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Resultsand implications. *Journal of Educational and Behavioral Statistics, 4*, 207-230.

Sabanathan, S., Wills, B., & Gladstone, M. (2015). Child development assessment tools in low- income and middle-income countries: how can we use them more appropriately? . *Arch Dis Child, 100*, 482 – 488. doi:doi:10.1136/archdischild-2014-308114

Saldana, J. (2013). *The coding manual for qualitative researchers* (2nd ed.). Los Angles: Sage.

Sampson, V., & Clark, D. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Science Education, 92*, 447 – 472.

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

Schommer, M. (1990). Effects of beliefs about the nature of knowledge on comprehension. *Journal of educational psychology, 82*(3), 498.

Senocak, E. (2009). Development of an instrument for assessing undergraduate science students' perceptions: The problem-based learning environment inventory. *Journal of Science Education and Technology, 18*(6), 560-569.

Shankoff, J., & Marshall, P. (2000). Chapter 2: Neurological Basis of Developmental Vulnerability. In: Handbook of Early Childhood Intervention, Cambridge: Cambridge University Press.

Shaw, S., & Crisp, V. (2012). An approach to validation: Developing and applying an approach for the validation of general qualifications. *Research Matters, Special*(3), 1-44.

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher, 36*(8), 477 – 481.

Sjøberg, S. (2018).The power and paradoxes of PISA: Should we sacrifice Inquiry-Based Science Education (IBSE) to climb on the Rankings? *NorDiNa, Nordic studies in science education 14(2), p186-202*

Smith, J. A. (2007). Hermeneutics, human sciences and health: Linking theory and practice. *International Journal of Qualitative Studies on Health and Well-being, 2*, 3–11.

Solomon, A. (2008). *Policy Formulation Curriculum Development and Implementation in Ethiopia*. Addis Ababa, Ethiopia: Addis Ababa University Press.

Stacey, O., Lazzari, G. D., Grayson, H., Griffin, H., Jones, E., Taylor, A., & Thomas, D. (2018). *The Globalization of Science Curricula*: Springer International Publishing.

Teshome, B. N. (2017). Classroom Participation and Development of Student Attitudes: A Study of Active Learning Practices in Ethiopian Primary Education.

Tiruneh, D. T., De Cock, M., Weldeslassie, A. G., Elen, J., & Janssen, R. (2017). Measuring critical thinking in physics: Development and validation of a critical thinking test in electricity and magnetism. *International Journal of Science and Mathematics Education, 15*(4), 663-682.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice,* 8-14.

Turiman, P., Omar, J., Daud, A., & Osman, K. (2012). Fostering the 21st century skills through scientific literacy and science process skills. *Procedia-Social and Behavioural Sciences,* 110-116.

Van der Graaf, J., Van de Sande, E., Gijsel, M., & Segers, E. (2019). A combined approach to strengthen children's scientific thinking: direct instruction on scientific reasoning and training of teacher's verbal support. *International Journal of Science Education, 41*(9), 1119-1138. Retrieved from https://doi.org/10.1080/09500693.2019.1594442.

Verspoor, A. M. (2008). *At the crossroads: choices for secondary education in Sub-Saharan Africa*: The World Bank.

Vieira, K., & De Queiroz, G. (2017). Hermeneutic Content Analysis: a method of textual analysis. *International Journal of Business Marketing and Management (IJBMM), 2*(8), 8-15.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching, 46.*(6), 716-730.

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

Zeineddin, A., & Abd-El-Khalick, F. (2010). Scientific reasoning and epistemological commitments: Coordination of theory and evidence among college science students. *Journal of Research in Science Teaching, 47*(9), 1064-1093.

Zhou, S., Han, J., Koenig, K., Raplinger, A., Pi, Y., Li, D., . . . Bao, L. (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.

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental review, 20*(1), 99-149.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental review, 27*(2), 172-223.

Appendix A: Final Scientific Reasoning Progress Tool (SRPT)

Question 1. Assume you are in a car traveling from Addis to Hosanna, covering 60 kilometers in one hour. Consider the car's start and finish points when answering the questions below.

1.1 Which physical quantity represents 60km?

- **a) <u>Distance</u>**
- b) Displacement
- c) Speed
- d) Velocity

1.3 How can you describe your state of motion during this journey from Addis to Hosanna?

- a) I am always in motion
- **b) <u>I am some times in motion and sometimes at rest</u>**
- c) I am always at rest
- d) I am neither at rest nor in motion

1.3 What evidence do you have to select one of the alternatives of question 1.2?

- a) It is because there is a change in position
- b) It is because the car that ran from Addis to Hossana
- **c) <u>It is because I am seated in the car but in motion relative to the earth</u>**
- d) It is because only the tire of the car that moves.

1.4 What can you conclude from the evidence you provided for question 1.3?

- a) A body is said to be in motion when there is change of position
- **b) <u>A body is said to be in motion when an object is in the state of moving</u>**
- c) A body is said to be in motion when it changes its position with time in a frame of reference
- d) A body is said to be in motion when the body feels state of change

Question 2: Suppose a ball is moving in a smooth horizontal frictionless surface.

2.1 Which physics law states that <u>if the ball is in motion with constant velocity it will continue its motion if no net external force is exerted</u>?

- **a) <u>Newton's 1st law</u>**
- b) Newton's 2nd law
- c) Newton's 3rd law
- d) There is no physics law related to the given statement.

2.3 What will happen to the motion of the ball **<u>if no net external force is exerted</u>** in question 2.1?

- a) It will slowdown and comes to rest.
- **b) <u>It will continue in motion with constant velocity</u>**
- c) It will speed up
- d) It is difficult to predict

2.3 Why do you select the above answer for question 2.2? It is because

- a) The ball will continue in motion until it loses its force within
- **b) <u>There is no change of state about the motion of the ball</u>**
- c) The ball continue in motion only if action and reaction forces are equal
- d) The ball continues in motion when only external force which is directly proportional to acceleration is exerted.

2.4 What can be concluded from question 2.3?

a) Moving objects always come to rest unless net external force is exerted on it.
b) Objects come to rest when they lose the force within them
c) <u>Objects tend to remain in its state without net external forces act on it.</u>
d) Objects move only when external force is exerted.

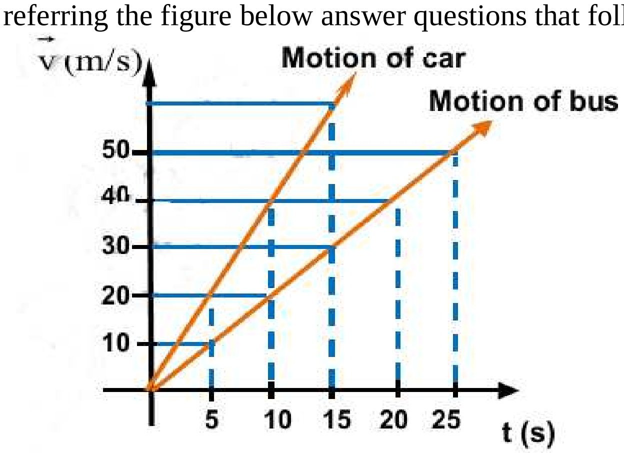
Question 3: Suppose a solid is heated at 50 ^{0}c.
3.1 What does 50 ^{0}c represents?
a) Heat
b) <u>Temperature</u>
c) Heat or temperature
d) There is no such a quantity

3.2 How can you describe about the state of molecules of the solid as heat increases?
a) <u>The molecules will move faster</u>
b) The molecules will move slower
c) The molecules will remain in constant state
d) The amount molecules will escape

3.3 Why do you select the above answer for question 3.2? It is because
a) <u>As heat increases speed of the molecules increases</u>
b) As heat increases speed of the molecules decreases
c) There is no relationship between heat and speed of the molecule.
d) As heat increases the mass of the molecules increases.

3.4 What can be concluded from questions 3.2 and 3.3?
a) Heat doesn't affect the motion of the molecules.
b) <u>Heating of substances will break the bonds between molecules and move apart.</u>
c) Heating of substances disturb motion of molecules
d) Heating substances increases the mass of molecules.

Question 4. In the figure below the graphs represent the motions of a bus and a car. By referring the figure below answer questions that follow



4.1 What is the speed of the bus at t = 5 seconds?
a) <u>10 m/s</u>
b) 15 m/s.
c) 20 m/s
d) 30 m/s

4.2 Which of the following statements correctly describes the relative speeds of the car and bus?
a) The bus is faster than the car
b) <u>The car is faster than the bus</u>

c) Both are moving at the same speed
d) It is difficult to predict

4.3 What is your reason for question number 4.2?

a) The bus has greater change of velocity than the car in time interval
b) <u>The car has greater change of velocity than the bus in time interval</u>
c) The graph of the car in the figure above is closer to the velocity axis
d) The graph of the bus in the figure above is closer to the time axis

4.4 What can you conclude from question 4.3?

a) The Car is always faster than bus
b) <u>The object with the greatest rate of change in velocity is the fastest</u>
c) Moving objects can never have equal speeds
d) There is not enough information given to draw conclusion.

Question 5: Suppose an apple hung on a tree.

5.1 What type of energy does the apple hold at the highest point?

a) <u>Potential energy</u>
b) Internal energy
c) Kinetic energy
d) Elastic potential energy

5.2 If the apple falls down, what will happen to the energy just before the apple hits the ground?

a) Kinetic energy will be converted to potential energy
b) The type of energy at the top will remain the same
c) <u>Potential energy will be converted to kinetic energy</u>
d) Some amount of energy will be destroyed

5.3 Why do you select the above answer for question 5.2? It is because

a) As the apple moves down its height increases but speed decreases
b) <u>As the apple moves down its height decreases but speed increases</u>
c) As the apple moves down both its height and speed decreases
d) As the apple moves down its height decreases but speed remain constant

5.4 What can be best concluded from question 5.3?

a) Kinetic energy depends both on mass and velocity
b) <u>Potential energy depends on height</u>
c) Potential energy depends both on mass and height
d) Kinetic energy depends on height.

Question 6: A teacher asked a group of six students to measure room temperature of a certain day and to report the result to the class. Assume that all of the students are using the same thermometer. The students' measurements are listed in the table below. Based on the table, answer questions 6.1-6.4.

6.1 Which one of the following is the repeated measurement from the table on the right?

Measured by	**Temperature** (^{o}C)
Student 1	25
Student 2	25.5
Student 3	26
Student 4	24
Student 5	26
Student 6	40

a) 25^{o} C
b) <u>26^{o} C</u>
c) 24^{o} C
d) 25.5^{o} C

6.2 Which student's measurement is important for reporting the result?

a) All the students measurements are important
b) **<u>Except for the student 6's, all of the students' measurements are important.</u>**
c) Only the student 6's measurement is important.
d) Only the measurement of students 3's and 5's are important

6.3 Why do you select the above answer for question 6.2? because

a) The highest measurement is more important.
b) All students use the same thermometer.
c) <u>Measurement of the student 6's is far from other measurements.</u>
d) Only the repeated measurements are important to make decision.

6.4 Which statement best concludes about the students' measurement?

a) There is no error in measurement
b) <u>All the measurements are not expected to be exactly the same</u>
c) The most repeatedly measured value is the correct measurement
d) The nearest value to the text book is the correct measurement

Question 8: An experiment is carried out by two students. To do so, they gather a variety of functional batteries, wires, and an ammeter. They measured the current values in the circuit after properly connecting it by altering the number of batteries for a single resistance. The following is a table of the students' measurements:

Voltage (v)	Resistance (Ω)	Current measured by first student (A)	Current measured by second student (A)
1.5	2	0.73	0.76
3	2	1.40	1.45
4.5	2	2.20	2.20
6	2	2.95	3.05

8.1 Which of the following measurements is **<u>not</u>** the value of current measured by the first student?

a) 1.40 A
b) 0.73 A
c) 2.20 A
d) <u>1.45 A</u>

8.2 In the above experiment what do the students want to know?

a) <u>They want to know the relationship between voltage and current</u>
b) They want to know how the resistance affects current.
c) They want to know the relationship between resistance and current
d) They want to know how the resistance affect voltage

8.3 Why do you select the above answer for the question number 8.2? It is because

a) The measurement is taken on voltage and current at constant resistance
b) <u>The measurement is taken on current by varying voltage at constant resistance</u>
c) The measurement is taken on voltage and resistance is calculated
d) The measurement is taken on voltage, current, and resistance.

8.5 What would you expect if the students use a 7.5 volt battery in the above experiment?

a) Student A and Student B will measure current that lies in the range between 2.95 and 3.05A.
b) Student A and Student B will measure current less than 2.95 A
c) Student A and Student B will measure current less than 3.05A
d) <u>Student A and Student B will measure current greater than 2.95A and 3.05A respectively.</u>

Question 9. A man pushes a body by 20N force and produces an acceleration of $2m/s^2$.

9.1 What are the given variables in question 9?

a) Force and velocity
b) Mass and acceleration
c) <u>Force and acceleration</u>
d) Force and speed

9.2 What force will produce the acceleration of $4m/s^2$ for constant mass in question 9?

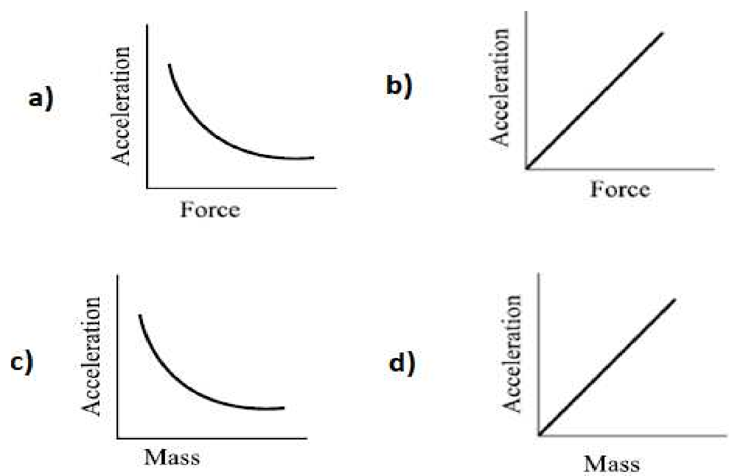
a) 10N
b) 20N

c) **40N**
d) 80N

9.3 Why do you select the above answer (to question 9.2)? It is because

a) **There is direct relationship between force and acceleration**
b) There is inverse relationship between mass and acceleration
c) To determine its acceleration mass should be given
d) The same force can produce different acceleration whether mass changes or remains constant.

9.4 Which one of the following graphs represents your answer to question number 9.3? **ANS: b**



Question10: A man lifts an object of 30 kg vertically to a height of 2 m (case 1), and an object of 20 kg vertically to a height of 3 m (case 2).

10.1 Which of the following variables *are* required to calculate gravitational potential energy?

e) Mass and velocity
f) Height and velocity
g) Velocity and gravitational acceleration.
h) **Mass and height**

10.2 In which cases work done by the man to lift the object is greater?

a) In the first case
b) The same work for both cases
c) In the second case
d) Impossible to predict

10.3 Why do you select the above answer for question 10.2? It is because

a) The man has lifted more mass in the first case
b) The man has lifted the object to a greater height in the second case
c) **Both masses possess the same amount of gravitational potential energy**
d) Two different situations can't be compared

10.4 Which statements below best conclude the situation in question 10?

a) Potential energy varies with mass of the object and acceleration due to gravity
b) **Potential energy varies with mass of the object and height the object has lifted**
c) Potential energy varies with height, mass of the object and acceleration due to gravity.

d) Potential energy varies with acceleration due to gravity and height a body has lifted.

Question11: Your neighbor needs a 2.0 -ohm resistor but he has two 4.0 -ohm resistors. He asks an electrician to help him how to produce a 2.0 -ohm resistor. To get a 2.0 -ohm resistor, the electrician tells him to put two 4.0-ohm resistors in parallel.

11.1 Which equation is helpful to find the total resistance of resistors connected in **parallel**? **Ans: a**

a) $\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2}$
b) $R = R_1 + R_2$
c) $R = \frac{1}{R_1} + \frac{1}{R_2}$
d) $\frac{1}{R} = R_1 + R_2$

11.3 What is your opinion about the electrician's advice?

d) I agree with the electrician
e) I disagree with the electrician
f) I agree to some extent.
d) I am not sure

11.3 Why do you agree, disagree, agree to some extent or not sure with the electrician in 11.2? It is because

a) In parallel connection, voltage is constant
b) The total resistance of two 4.0 -ohm resistors in parallel gives a 2.0 Ω resistor
c) The total resistance of two 4.0 -ohm resistors in series gives a 2.0 -ohm resistor
d) Resistance cannot be added, rather made with fixed values.

11.4 What can you conclude from your answer for the question 11.3?

a) Total resistance can be greater than individual resistances
b) Total resistance can be less than individual resistances
c) Resistances can be added by both series and parallel connections.
d) The electrician knows better than others

Appendix B: Fit statistics for final version of SRPT

Item No	Model Measure	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	Pt. Corr.
Item 1.1	-1.44	1.08	1.2	1.15	1.6	.26
Item 1.2	-.13	1.04	.9	1.06	.8	.31
Item 1.3	.57	.99	-.2	1.05	.5	.32
Item 1.4	.58	1.00	.0	.98	-.1	.32
Item 2.1	-1.10	1.03	.6	1.00	.1	.35
Item 2.2	-.33	1.05	1.0	1.08	1.3	.30
Item 2.3	.81	1.03	.4	1.09	.7	.25
Item 2.4	1.19	.94	-.5	.79	-1.3	.36
Item 3.1	-1.77	.95	-.6	.92	-.6	.41
Item 3.2	-.91	.81	-3.6	.78	-3.4	.58
Item 3.3	-.32	.90	-2.1	.92	-1.3	.47
Item 3.4	.58	1.08	1.2	1.09	.8	.24
Item 4.1	-1.52	.97	-.4	.94	-.6	.40
Item 4.2	-.64	.97	-.6	.98	-.3	.40
Item 4.3	-.29	.85	-3.3	.83	-2.8	.53
Item 4.4	.45	.94	-1.0	.86	-1.4	.41
Item 5.1	-.88	1.10	1.8	1.10	1.5	.27
Item 5.2	-.35	1.08	1.7	1.09	1.4	.28
Item 5.3	.33	1.03	.5	1.01	.1	.31
Item 5.4	1.79	1.00	.1	.86	-.5	.24
Item 6.1	-1.37	.96	-.5	.92	-1.0	.42
Item 6.2	-.03	1.06	1.3	1.12	1.6	.28
Item 6.3	.69	1.05	.7	1.05	.5	.26
Item 6.4	1.36	.96	-.3	.79	-1.1	.33
Item 8.1	-1.23	.96	-.6	.97	-.3	.41
Item 8.2	.34	1.02	.4	1.02	.3	.31
Item 8.3	.94	1.00	.1	.92	-.5	.30
Item 8.4	1.44	1.04	.4	1.14	.7	.18
Item 9.1	-1.16	.91	-1.5	.87	-1.8	.48
Item 9.2	-.49	1.10	1.9	1.11	1.7	.27
Item 9.3	.26	.96	-.6	.87	-1.5	.40
Item 9.4	.67	1.01	.2	1.09	.8	.29
Item 10.1	-.67	1.09	1.7	1.11	1.8	.28
Item 10.2	.09	1.01	.3	1.04	.5	.33
Item 10.3	.37	.99	-.1	.96	-.4	.34
Item 10.4	1.00	1.07	.8	1.06	.4	.21
Item 11.1	-.91	1.04	.7	1.02	.4	.34
Item 11.2	-.24	1.05	1.0	1.07	1.1	.31
Item 11.3	.63	.95	-.7	.91	-.7	.38
Item 11.4	1.68	.93	-.5	.73	-1.3	.33

Appendix C: Item category entry order for the final version of SRPT

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	%	AVERAGE MEASURE	S.E. MEAN	OUTF MNSQ	PTMEA CORR.	ITEM
1	.	***	2	1#	-.05	.09		.05	ITEM1.1
	0	0	73	30	-.84	.11	1.2	-.26	
	1	1	167	70	-.35	.06	1.1	.26	
2	.	***	11	5#	-.46	.30		.01	ITEM1.2
	0	0	133	58	-.72	.07	1.0	-.31	
	1	1	98	42	-.19	.08	1.1	.31	
3	.	***	4	2#	-.09	.19		.06	ITEM1.3
	0	0	170	71	-.67	.06	1.0	-.32	
	1	1	68	29	-.07	.09	1.1	.32	
4	.	***	2	1#	.21	.31		.08	ITEM1.4
	0	0	172	72	-.67	.06	1.0	-.32	
	1	1	68	28	-.06	.08	1.0	.32	
5	.	***	3	1#	-.98	.61		-.06	ITEM2.1
	0	0	88	37	-.87	.09	1.0	-.35	
	1	1	151	63	-.26	.06	1.0	.35	
6	.	***	13	5#	-.70	.28		-.06	ITEM2.2
	0	0	121	53	-.72	.07	1.0	-.30	
	1	1	108	47	-.21	.08	1.1	.30	
7	.	***	11	5#	-.22	.25		.07	ITEM2.3
	0	0	175	76	-.63	.06	1.0	-.25	
	1	1	56	24	-.13	.10	1.1	.25	
8	.	***	13	5#	-.10	.12		.11	ITEM2.4
	0	0	187	82	-.66	.06	1.0	-.36	
	1	1	42	18	.14	.09	.8	.36	
9	.	***	5	2#	-.16	.23		.06	ITEM3.1
	0	0	59	25	-1.11	.11	.9	-.41	
	1	1	178	75	-.30	.06	1.0	.41	
10	.	***	13	5#	-.35	.17		.04	ITEM3.2

10	.	***	13	5#	-.35	.17		.04	ITEM3.2
	0	0	94	41	-1.10	.08	.8	-.58	
	1	1	135	59	-.08	.06	.8	.58	
11	.	***	7	3#	-.18	.20		.06	ITEM3.3
	0	0	126	54	-.88	.07	.9	-.47	
	1	1	109	46	-.07	.07	1.0	.47	
12	.	***	11	5#	-.19	.16		.08	ITEM3.4
	0	0	166	72	-.63	.07	1.1	-.24	
	1	1	65	28	-.18	.08	1.1	.24	
13	.	***	4	2#	-.12	.26		.06	ITEM4.1
	0	0	69	29	-1.03	.10	.9	-.40	
	1	1	169	71	-.28	.06	1.0	.40	
14	.	***	16	7#	.05	.19		.17	ITEM4.2
	0	0	107	47	-.89	.08	.9	-.40	
	1	1	119	53	-.21	.07	1.0	.40	
15	.	***	8	3#	-.21	.16		.06	ITEM4.3
	0	0	127	54	-.92	.07	.9	-.53	
	1	1	107	46	-.01	.06	.8	.53	
16	.	***	7	3#	.03	.18		.11	ITEM4.4
	0	0	163	69	-.74	.07	1.0	-.41	
	1	1	72	31	.02	.07	.8	.41	
17	.	***	26	11#	-.19	.13		.12	ITEM5.1
	0	0	91	42	-.81	.10	1.1	-.27	
	1	1	125	58	-.33	.07	1.1	.27	
18	.	***	9	4#	-.75	.27		-.06	ITEM5.2
	0	0	122	52	-.71	.08	1.1	-.28	
	1	1	111	48	-.23	.07	1.1	.28	
19	.	***	14	6#	-.23	.15		.08	ITEMN5.3

20	.	***	13	5#	-.18	.20		.09	ITEM5.4
	0	0	203	89	-.58	.06	1.0	-.24	
	1	1	26	11	.05	.10	.8	.24	
21	.	***	10	4#	.52	.19		.25	ITEM6.1
	0	0	75	32	-1.04	.09	.9	-.42	
	1	1	157	68	-.30	.06	1.0	.42	
22	.	***	2	1#	-.47	.36		.00	ITEM6.2
	0	0	143	60	-.69	.07	1.0	-.28	
	1	1	97	40	-.21	.08	1.2	.28	
23	.	***	3	1#	.14	.24		.08	ITEM6.3
	0	0	176	74	-.63	.07	1.0	-.26	
	1	1	63	26	-.14	.09	1.1	.26	
24	.	***	15	6#	-.44	.17		.02	ITEM6.4
	0	0	190	84	-.62	.06	1.0	-.33	
	1	1	37	16	.14	.09	.8	.33	
25	.	***	11	5#	-.63	.28		-.03	ITEM8.1
	0	0	79	34	-.96	.10	1.0	-.41	
	1	1	152	66	-.24	.06	.9	.41	
26	.	***	11	5#	-.39	.24		.03	ITEM8.2
	0	0	155	67	-.68	.07	1.0	-.31	
	1	1	76	33	-.12	.08	1.0	.31	
27	.	***	11	5#	-.17	.20		.08	ITEM8.3
	0	0	180	78	-.64	.06	1.0	-.30	
	1	1	51	22	-.03	.08	.9	.30	
28	.	***	15	6#	-.08	.18		.12	ITEM8.4
	0	0	193	85	-.58	.06	1.0	-.18	
	1	1	34	15	-.15	.13	1.2	.18	
29	.	***	6	2#	-.65	.42		-.03	ITEM9.1

30	.	***	20	8#	-.72	.14		-.08	ITEM9.2
	0	0	109	49	-.71	.09	1.2	-.27	
	1	1	113	51	-.24	.07	1.1	.27	
31	.	***	17	7#	.00	.14		.16	ITEM9.3
	0	0	149	66	-.78	.07	1.0	-.40	
	1	1	76	34	-.05	.06	.8	.40	
32	.	***	24	10#	-.39	.15		.04	ITEM9.4
	0	0	160	73	-.65	.07	1.0	-.29	
	1	1	58	27	-.09	.10	1.1	.29	
33	.	***	10	4#	-.52	.27		-.01	ITEM10.1
	0	0	106	46	-.75	.08	1.1	-.28	
	1	1	126	54	-.28	.07	1.2	.28	
34	.	***	8	3#	-.48	.29		.00	ITEM10.2
	0	0	145	62	-.71	.07	1.0	-.33	
	1	1	89	38	-.13	.08	1.1	.33	
35	.	***	13	5#	-.08	.26		.12	ITEM10.3
	0	0	156	68	-.71	.07	1.0	-.34	
	1	1	73	32	-.10	.08	.9	.34	
36	.	***	21	9#	-.08	.16		.15	ITEM10.4
	0	0	175	79	-.62	.07	1.1	-.21	
	1	1	46	21	-.18	.09	1.1	.21	
37	.	***	10	4#	-.79	.27		-.07	ITEM11.1
	0	0	94	41	-.83	.09	1.0	-.34	
	1	1	138	59	-.24	.06	1.0	.34	
38	.	***	17	7#	-.23	.20		.09	ITEM11.2
	0	0	125	56	-.75	.08	1.0	-.31	
	1	1	100	44	-.22	.08	1.1	.31	
39	.	***	33	14#	-.38	.13		.05	ITEM11.3

Appendix D: Interview schedule by the researcher (initial)

1. What is physics for you? How do you perceive/view physics knowledge before and after learning the concepts of physics?
2. How do you view physics facts and concepts?
3. Is there any relationship between physics concepts and day to day activities? How can you relate physics facts, concepts with what you know in regular basis? (3)
4. Does the construction/development of scientific knowledge require evidences? If so, what are sources of evidences?
5. Can we arrive on a certain conclusion based on the evidences? How can one draw a scientific conclusion?

Appendix E: : Validated Interview Schedule for the Students

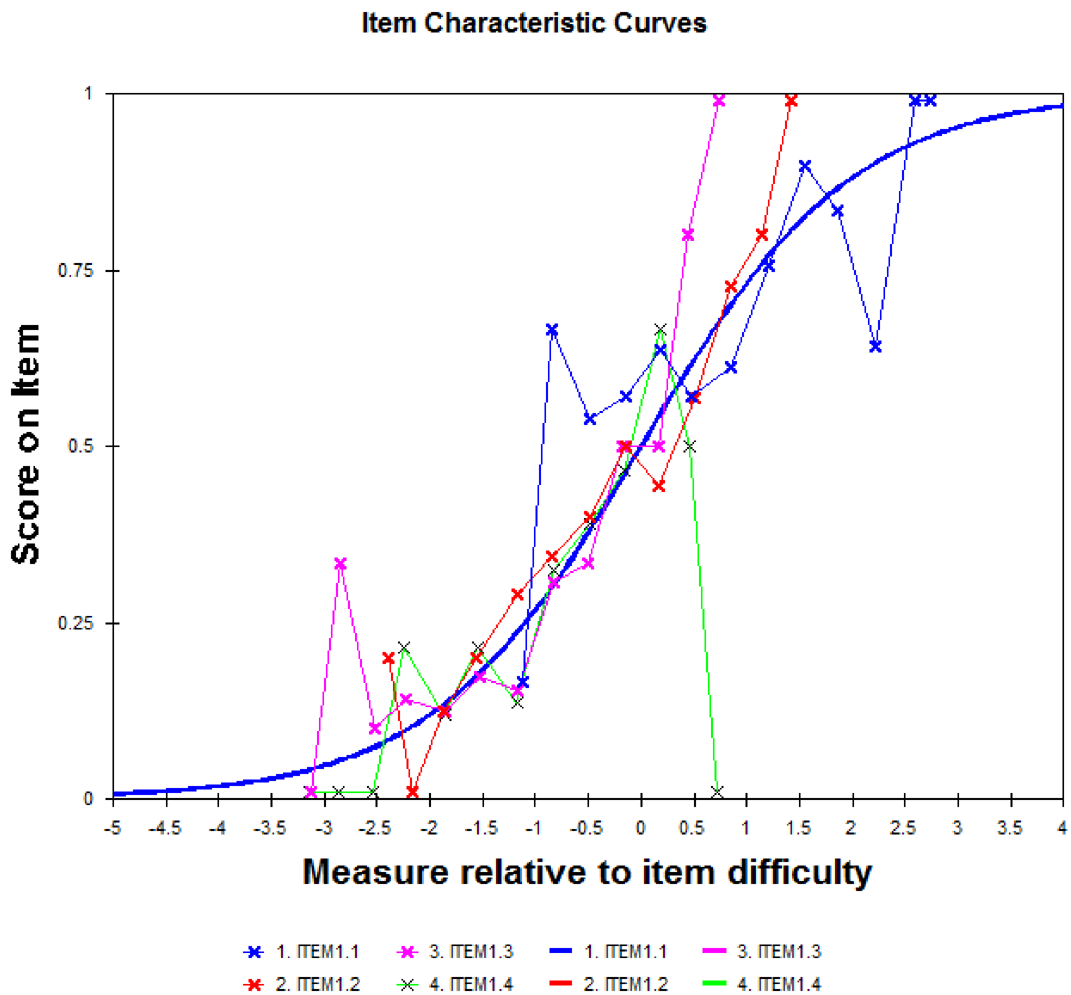
1. How do you perceive/view physics/scientific knowledge before and after learning the concepts of physics?
2. How do you view physics facts and concepts? *(Probing question: Do you think the definitions, constants, formulas…in the textbook are truths to be grasped or something needs to be checked, proved, and supported with evidences? Give an example from physics concept.)*
3. How do you view scientific explanations? *(Probing questions: how scientific explanations are constructed? Is there any relationship between scientific explanations and day to day experiences? Give an example from physics concept.)*
4. Does the construction/development of scientific knowledge require evidences? *If so, what are the sources of evidences? (give an example from physics concept)*
5. Can we arrive on a certain conclusion based on the evidences? *(Probing question: How can one draw a scientific conclusion?)*
6. How can you develop/construct your understanding of physics/scientific knowledge from the low to high level of understanding? (give an example from physics concept)

Appendix F: Questions Used in the Focus Group Discussion

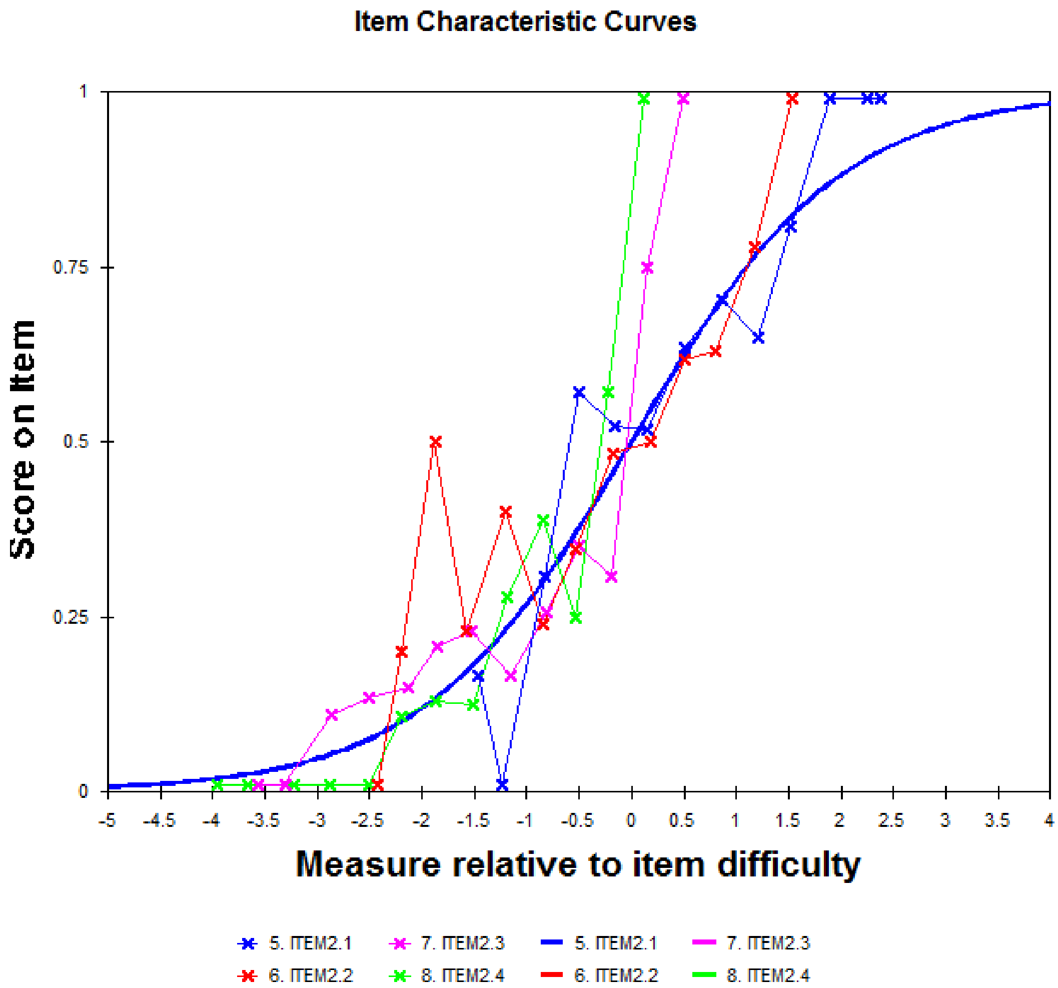
1. How do students' perceive/view physics/scientific knowledge?
2. How students view physics facts and concepts?
3. How do students view physics/scientific explanations?
4. Do students generate their own evidences for their views? *If so, what type of reasoning?*
5. Do students able to draw conclusions? If so, what kind of conclusion?
6. How can you put students' progress of reasoning from low level to high level of reasoning from your experience (if possible)?

Appendix G: ICC curves for SRPT items

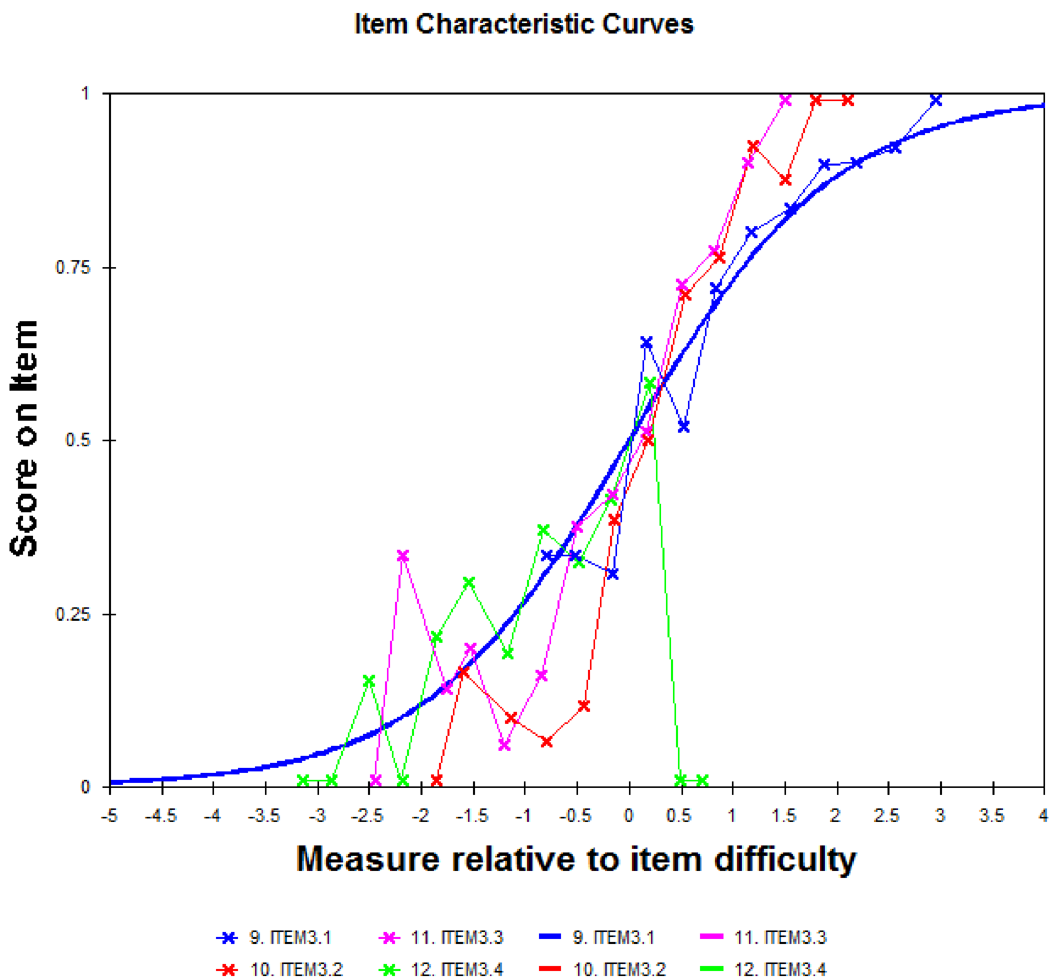
a) ICC curve foitems 1.1, 1.2, 1.3 and 1.4 in terms of score and its difficulty



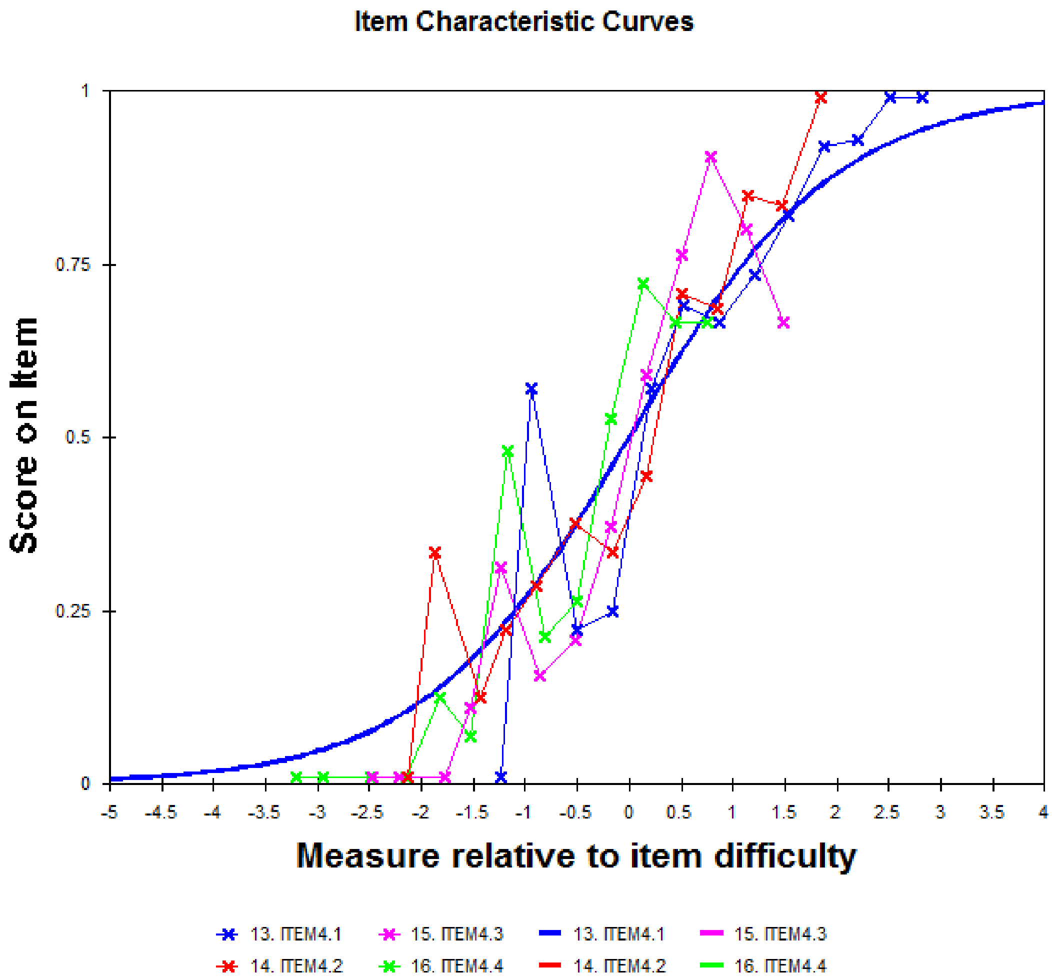
b) ICC curve fo items 2.1, 2.2, 2.3 and 2.4 in terms of score and its difficulty



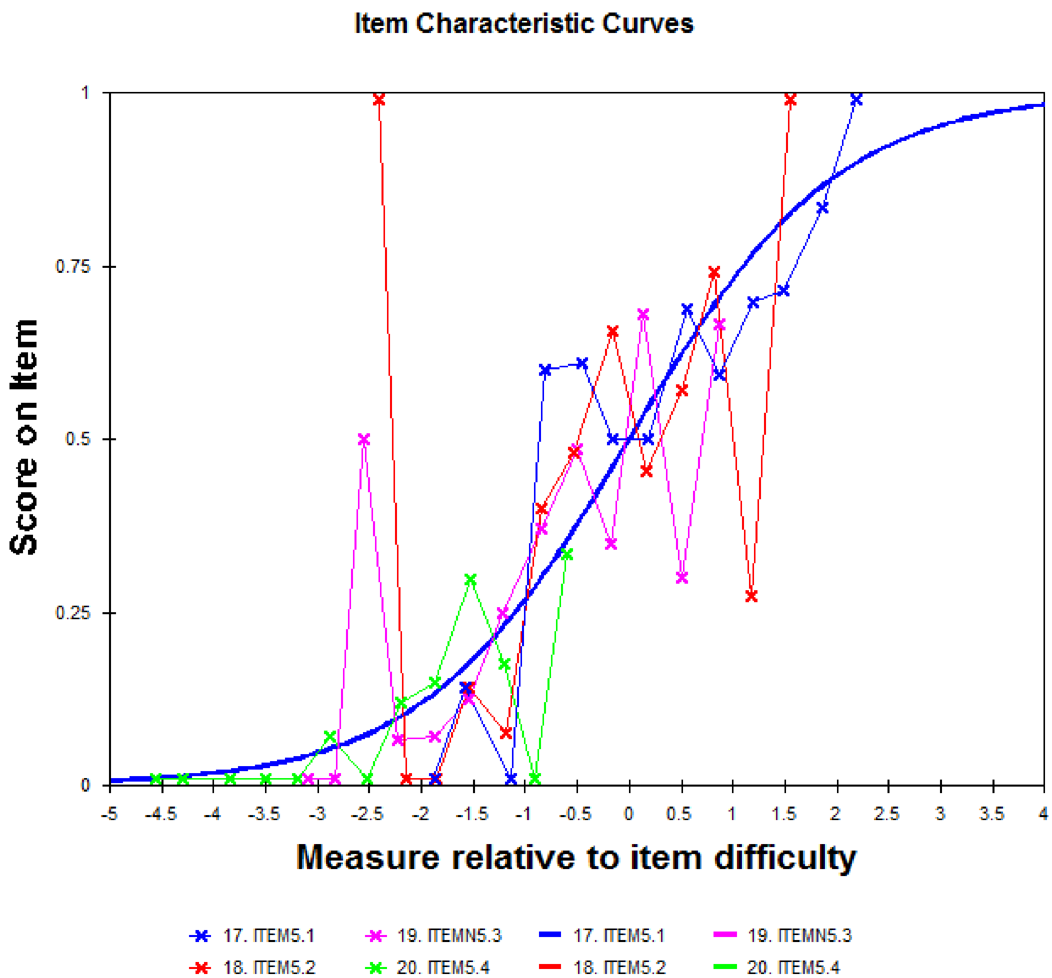
c) ICC curve for items 3.1, 3.2, 3.3 and 4.4 in terms of score and its difficulty



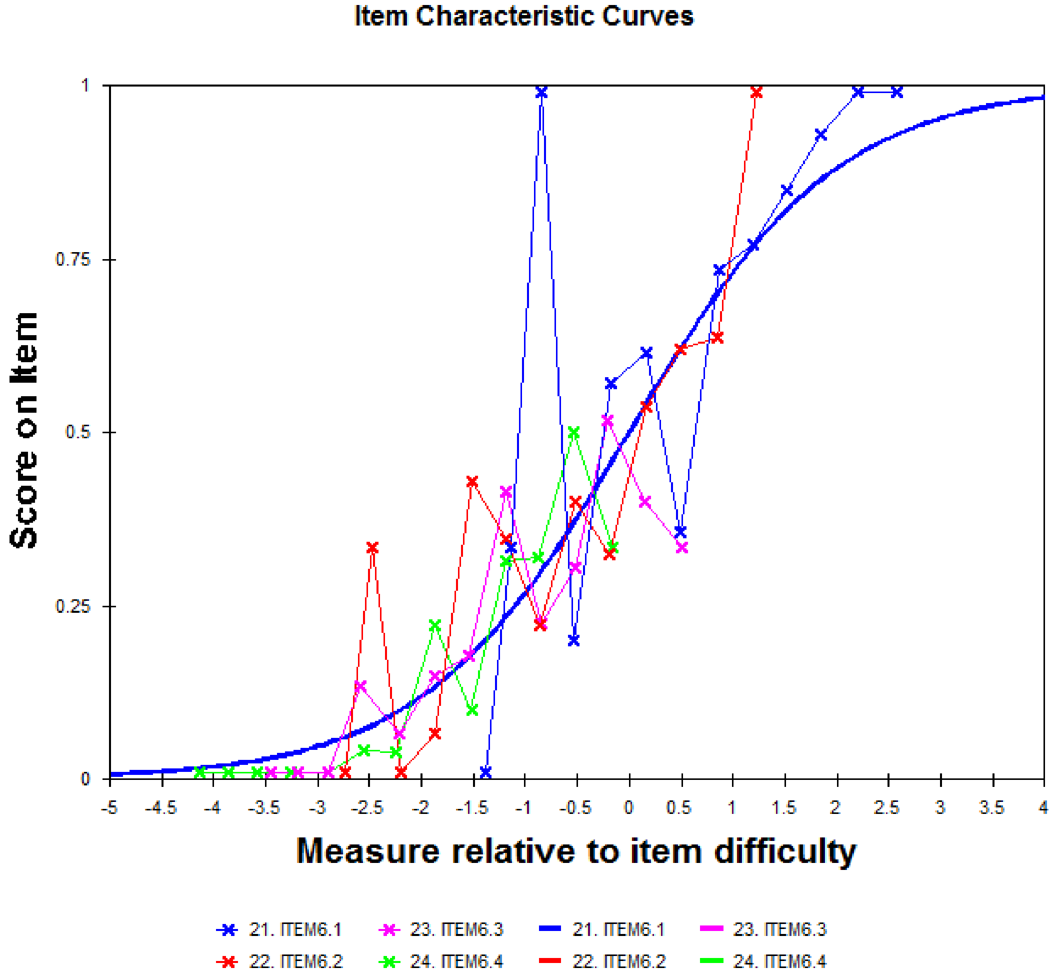
d) ICC curve for items 4.1, 4.2, 4.3 and 4.4 in terms of score and its difficulty



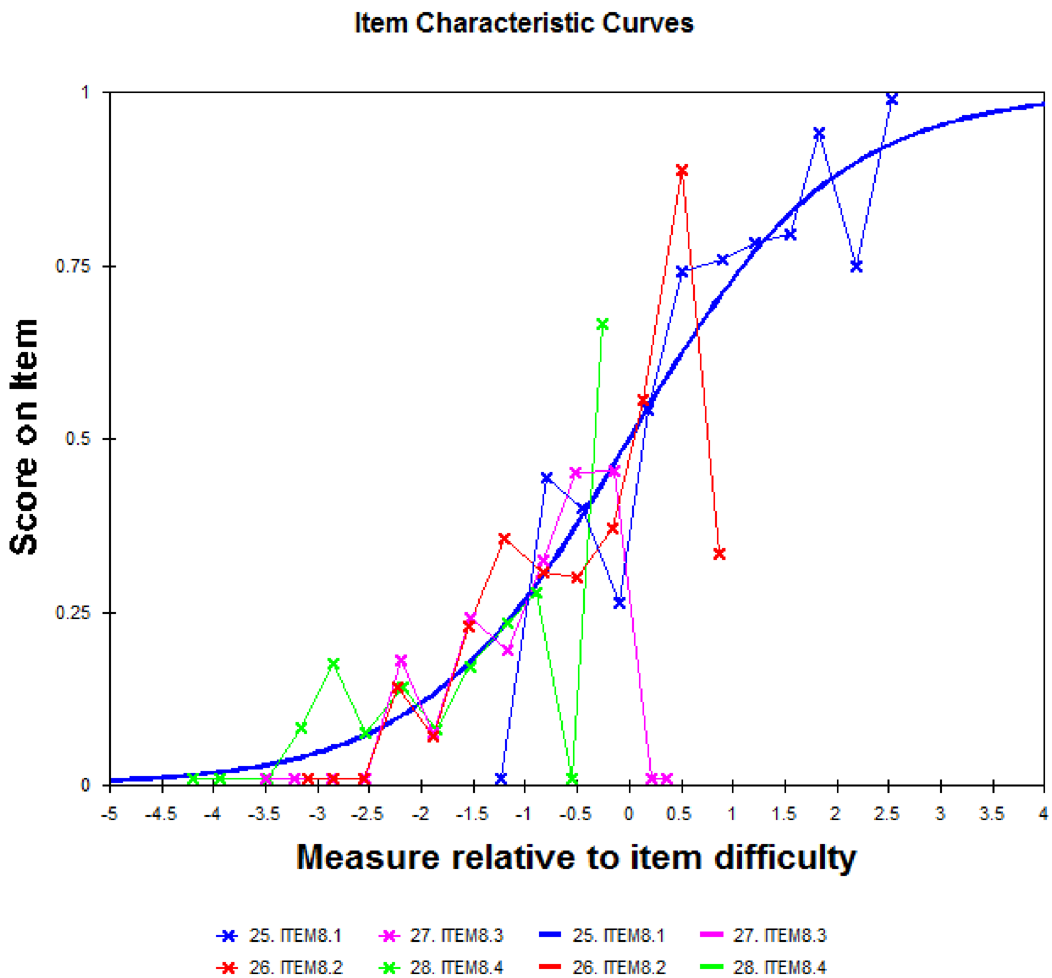
e) ICC curve for items 5.1, 5.2, 5.3 and 5.4 in terms of score and its difficulty



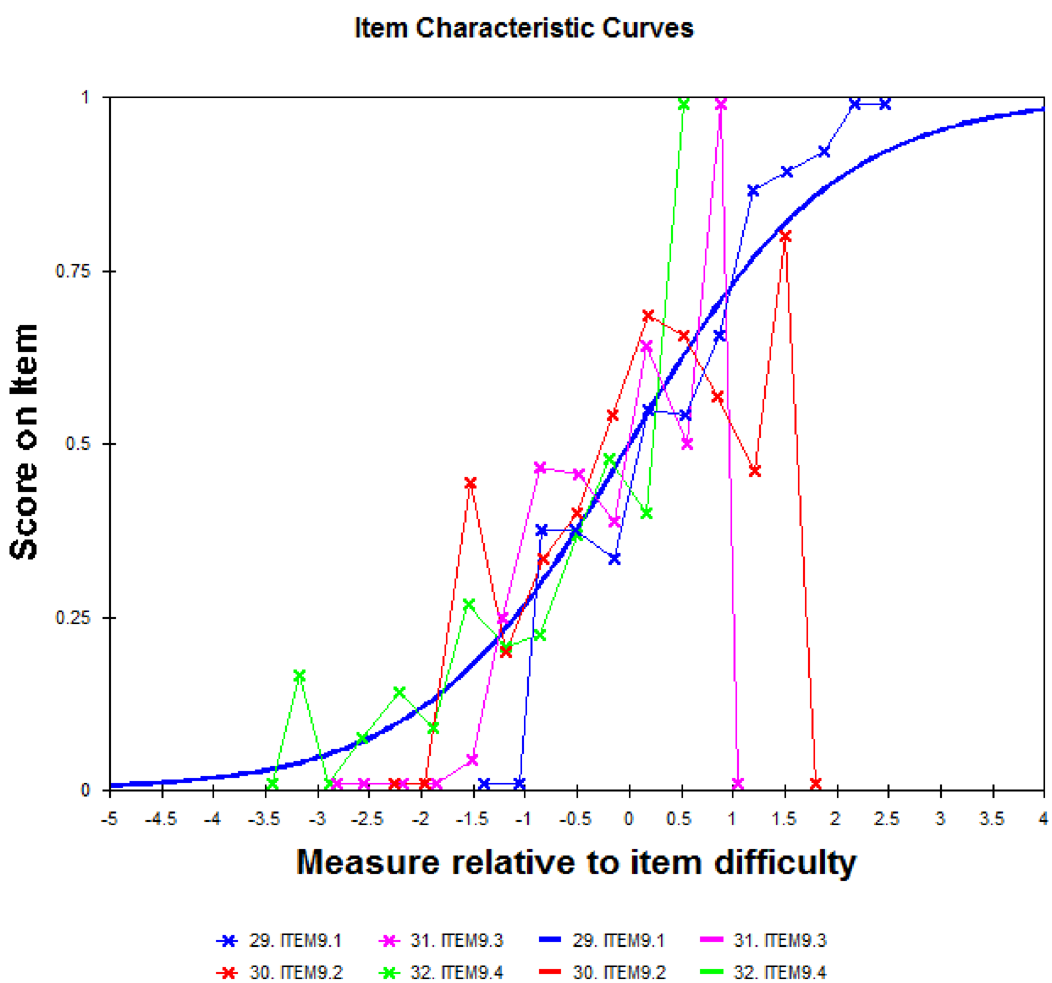
f) ICC curve for items 6.1, 6.2, 6.3 and 6.4 in terms of score and its difficulty



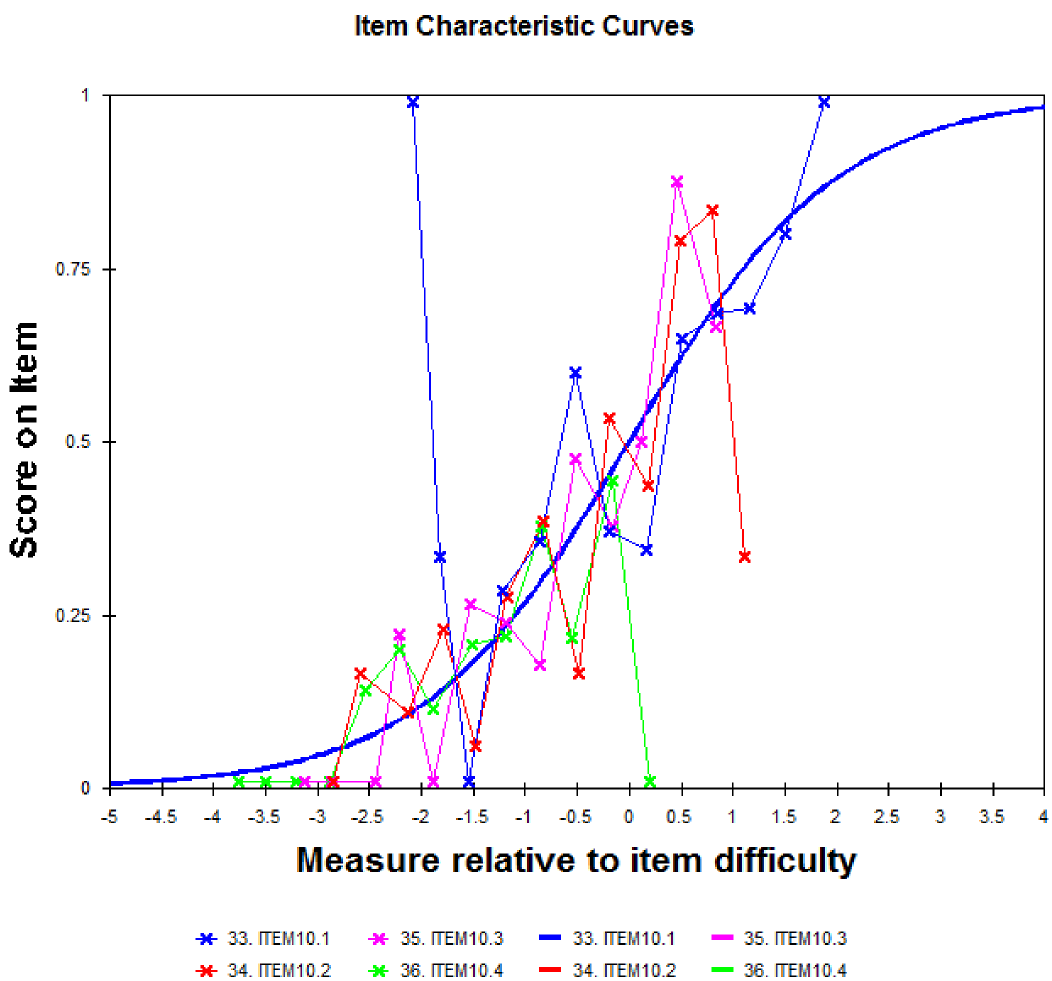
g) ICC curve for items 8.1, 8.2, 8.3 and 8.4 in terms of score and its difficulty



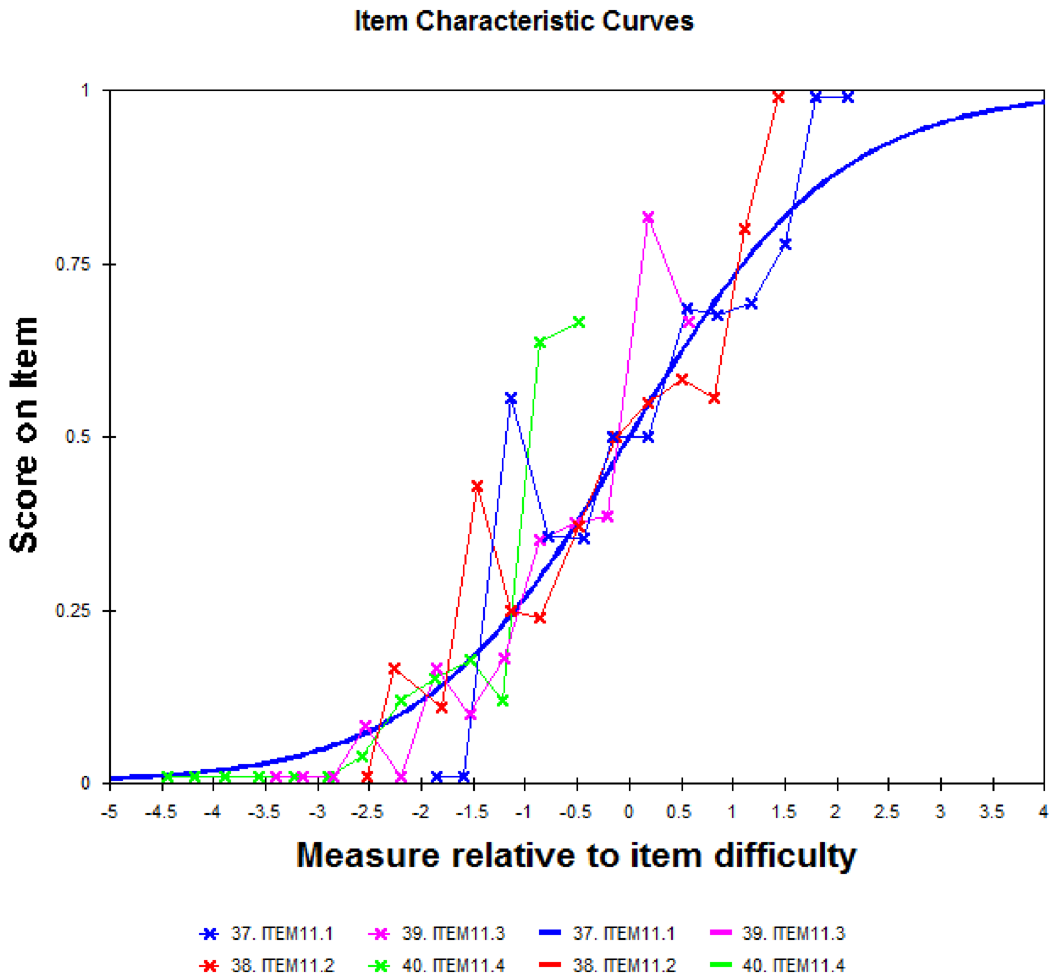
h) ICC curve for items 9.1, 9.2, 9.3 and 9.4 in terms of score and its difficulty



i) ICC curve for items 10.1, 10.2, 10.3 and 10.4 in terms of score and its difficulty



j) ICC curve for items 11.1, 11.2, 11.3 and 11.4 in terms of score and its difficulty



Item Characteristic Curves
Score on Item
1
0.75
0.5
0.25
0
-5
-4.5
-4
-3.5
-3
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
3.5
4
Measure relative to item difficulty
37. ITEM11.1
39. ITEM11.3
37. ITEM11.1
39. ITEM11.3
38. ITEM11.2
40. ITEM11.4
38. ITEM11.2
40. ITEM11.4

Appendix H: Consent letter to the participating teachers

ADDIS ABABA UNIVERSITY

COLLEGE OF EDUCATIONAL AND BEHAVIORAL STUDIES

Informed Consent for the Teachers

Dear Sir/Madam,

The following information is provided to help you decide whether you wish to participate in the present study. You should be aware that you are free to decide not to participate or to withdraw at any time.

I am conducting this research for the fulfilment of a PhD degree in Physics Education. This interview/discussion was aimed to obtain information about grade 8 students' which can be used to develop assessmet tool and explore stduents' pattern of reasoning at upper primary school in physics.

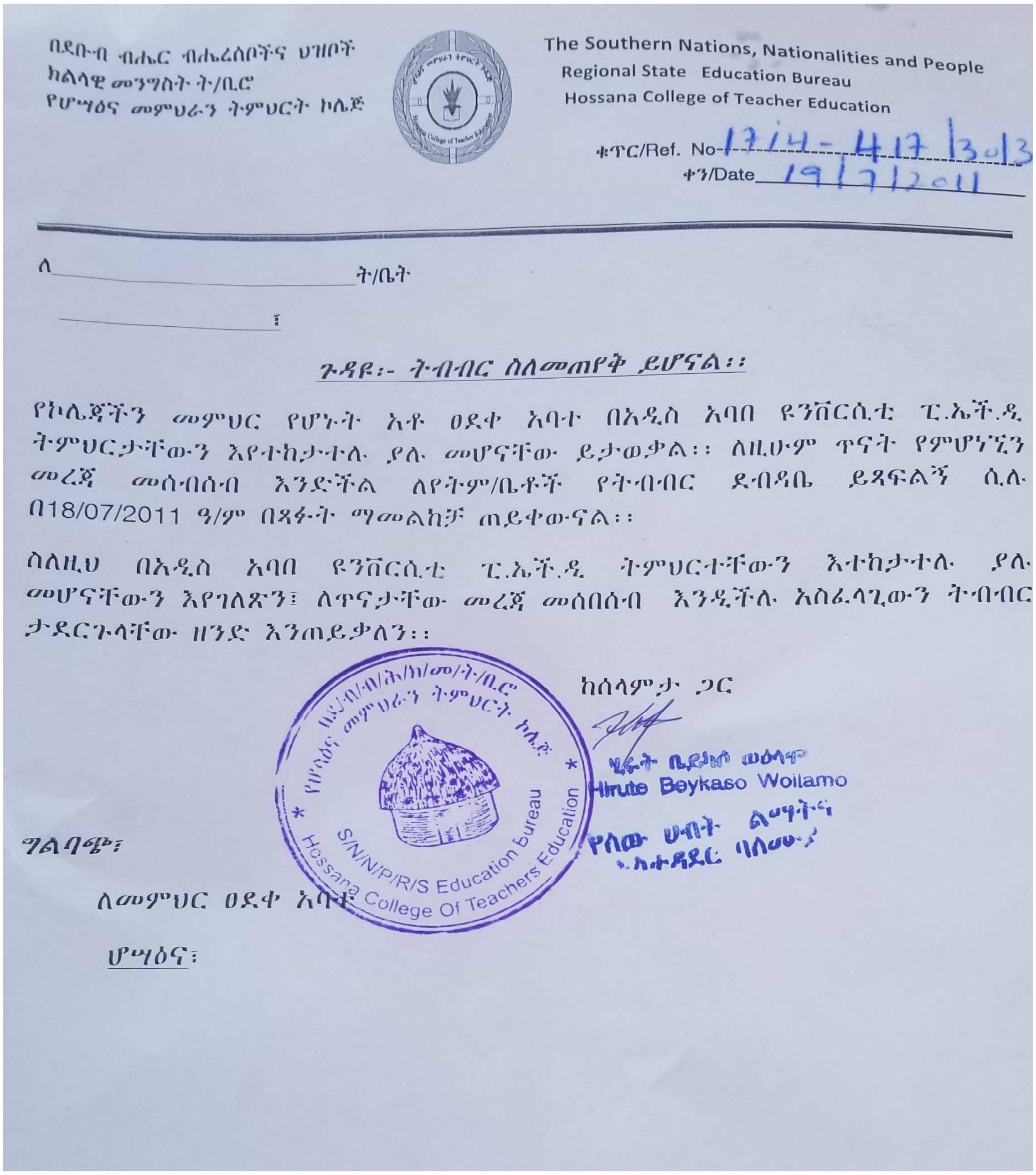
Thank you for your willingness to participate in this discussion. I would like to hear your ideas and opinions. I want to assure you that your responses to the questions will be kept confidential and your name will not be associated with the research findings in any way. Do not hesitate to ask questions about the study before participating or during the study. I would be happy to share the findings with you after the research is completed.

Please sign this consent form. You are signing it with full knowledge of the nature and purpose of the procedures. A copy of this form will be given to you to keep.

Name ----------------------------

Signature ------------------------------

Appendix I: Consent letter to the participating schools



በደቡብ ብሔር ብሔረሰቦችና ህዝቦች
ክልላዊ መንግስት ት/ቢሮ
የሆሳዕና መምህራን ትምህርት ኮሌጅ

The Southern Nations, Nationalities and People
Regional State Education Bureau
Hossana College of Teacher Education

ቁጥር/Ref. No- 1714-417/3013
ቀን/Date 19/7/2011

ለ ______________ ት/ቤት

______________ ፤

ጉዳዩ፡- ትብብር ስለመጠየቅ ይሆናል፡፡

የኮሌጃችን መምህር የሆኑት አቶ ፀደቀ አባተ በአዲስ አባበ ዩንቨርሲቲ ፒ.ኤች.ዲ ትምህርታቸውን እየተከታተሉ ያሉ መሆናቸው ይታወቃል፡፡ ለዚሁም ጥናት የምሆነኘን መረጃ መሰብሰብ እንድችል ለየትም/ቤቶች የትብብር ደብዳቤ ይጻፍልኝሲሉ በ18/07/2011 ዓ/ም በጻፉት ማመልከቻ ጠይቀውናል፡፡

ስለዚህ በአዲስ አባበ ዩንቨርሲቲ ፒ.ኤች.ዲ ትምህርተቸውን እተከታተሉ ያሉ መሆናቸውን እየገለጽን፤ ለጥናታቸው መረጃ መሰበሰብ እንዲችሉ አስፈላጊውን ትብበር ታደርጉላቸው ዘንድ እንጠይቃለን፡፡

ከሰላምታ ጋር

ሂሩት ቤይካሶ ወሰላሞ
Hirute Beykaso Woilamo

ግልባጭ፤

ለመምህር ፀደቀ አባተ

ሆሣዕና፤

Declaration

I, the undersigned candidate declare that this dissertation entitled: *Development of Tools and Assessment of Students' Pattern of Reasoning at Upper Primary School of Hadiyya Zone, Ethiopia* is an original research I conducted under the supervision of *Prof. Carl Angell and Dr. Kassa Michael*. The full or any part of this dissertation has not been submitted for any degree or examination in this or any other university and therefore, it is not by any means a replication of a work already done by no one. Further, I declare that all the sources I have used or quoted have been indicated and acknowledged as complete references in this dissertation.

Candidate name: ___ Tsedeke Abate ____________________________

Id Number: _________________________

Signature __________________________ Date: __________________________

I (we), Tsedeke Abate's dissertation supervisors testify that this dissertation has been developed under my (our) supervision and I (we) approved it to be submitted to the department for evaluation as a fulfillment of the requirements for the degree of Doctor of Philosophy.

Main Supervisor:

Name: **Professor Carl Angell_**____________________________

Signature: ____________________________ Date: ________________________

Co-Supervisor:

Name: **Dr. Kasaa Michael**____

Signature: ____________________________ Date: ________________________