MODESTUM
OPEN ACCESS

EURASIA Journal of Mathematics, Science and Technology Education, 2020, 16(12), em1927
ISSN:1305-8223 (online)
Research Paper
https://doi.org/10.29333/ejmste/9353

Assessment of Scientific Reasoning: Development and Validation of Scientific Reasoning Assessment Tool

Tsedeke Abate [1,2*], Kassa Michael [2], Carl Angell [3]

[1] Department of Science and Mathematics Education, Addis Ababa University, P. O. Box 1176, Addis Ababa, ETHIOPIA
[2] Department of Natural Sciences, Hossana College of Teachers Education, P.O. Box 94, Hossana, ETHIOPIA
[3] Department of Physics, University of Oslo, NORWAY

Received 2 August 2020 ▪ Accepted 16 October 2020

Abstract

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. This study aimed to develop and validate Scientific Reasoning Progress Tool (SRPT) that measures students' reasoning abilities. In this study, 40 items were developed, pilot-tested, and administered to 242 students from grade eight. The SRPT was a valid and reliable instrument. It was also found that the reasoning ability of grade 8 students' is limited to the lower levels of reasoning. It is recommended that further study is essential through the adoption of the framework and the design to develop additional instruments and investigation of the progression of students' scientific reasoning ability.

Keywords: Rasch analysis, reasoning progress, scientific reasoning, styles of scientific reasoning

INTRODUCTION

These days, science education is believed to be a means to tackle the challenges of the 21st century such as poverty in sub-Saharan countries (Josh & Verspoor, 2013), and poor quality of life and other societal problems (Zhou, et al., 2016). To solve such challenges of the 21st century, science education needs to help students to develop the skills needed for the Century such as the ability to solve problems, evaluate information, collaborate with others effectively, work with a variety of new technology, critical thinking, reasoning, and develop new ideas and products (Dole, Bloom, & Kowalske, 2016; Lamb, Jackson, & Rumberger, 2015; Nagaoka, Farrington, Ehrlich, & Heath, 2015). Science education is required to play a paramount role in equipping students with the skills that enable them to be competitive in the era of globalization, promote a rational culture, and make proper decisions.

Failure of science education for the needs of the 21st century is, however, to some extent attributed to its inability to incorporate scientific reasoning as a good model in the school instruction (Osborne, 2013). Studies indicate that scientific reasoning ability promotes students' skills in solving real-life problems (Han, 2013). It is considered a better predictor of success in science education (Osborne, 2013), and has an impact on students' long-term academic achievement (Bao et al., 2009). Scientific reasoning ability has also recently gained crucial significance in STEM subjects (Kind & Osborne, 2017; Optiz, Heene, & Fisher, 2017). To promote and establish such ability in the science classroom, it is required to assess students' pattern of scientific reasoning and develop valid assessment tools that measure and evaluate their ability of scientific reasoning.

However, 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 (Opitz et al., 2017) without emphasizing the content knowledge.

✉ tsedaba@yahoo.com **(*Correspondence)** ✉ ftkassa2010@yahoo.com ✉ carl.angell@fys.uio.no

Contribution to the literature

- There is a need for context-based, construct-based, and domain-specific assessment tools that can measure current students' reasoning abilities in low-income nations such as countries in sub-Saharan Africa.
- There is a need to develop valid and reliable assessment tools that measures students' reasoning ability by considering the context of the students in terms of the curricula and their learning.
- There is a need to assess students' ability and level of scientific reasoning in low-income nations with valid and reliable tools.

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. According to OECD (2016) content knowledge involves knowledge of the facts, concepts, ideas, and theories about the natural world and the explanations. Procedural knowledge involves the procedures and associated constructs that science uses to establish its claims to know. Procedural knowledge is related to scientific evidence to accept or reject a claim (Gott, Duggan, & Roberts, 2008). Epistemic knowledge is about the epistemic constructs and values and how these are used to justify science's claims to know. Epistemic knowledge involves claims, explanations, evidence, hypotheses, models, and theories to establish scientific knowledge (OECD, 2016).

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, Rafanelli & Kind, 2018). Kind and Osborne (2017) argued that giving much focus on reasoning skills alone would not help to bring 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, 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.

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 which demand critical decisions and judgments but without considering evidence-based assessments (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 (Verspoor, 2008). This shift of educational system seems to match the students' learning with international trends. But, Ethiopia offers compartmentalized science subjects starting from grade 7 that seeks the development of subject specific reasoning and problem solving.

Despite the curricular organization, the current classroom is dominated by lower-level cognitive demands globally (Osborne, 2013) and particularly in low-income nations (Joshi & Verspoor, 2013). In many sub-Saharan African countries, current classroom assessments are limited in measuring students' low-level understanding such as recalling, memorization of facts, and ability to use algorithms (Joshi & Verspoor, 2013; Teshome, 2017; Verspoor, 2008). To solve such problems several international assessments and studies are engaged in developing scientific reasoning assessments and students' level of reasoning. Some of them are Lawson's Classroom Test of Scientific Reasoning (CTSR), PISA, TIMSS, PIRLS, and NAEP. These assessments are dominant, well-established for a long

time, and items from these studies are commonly used in scientific research (Bao et al., 2009; DeBoer, 2011).

The assessment projects such as PISA, TIMSS, and NAEP have developed and administered tests to assess students, in science and mathematics, at various grade levels in many countries around the globe by considering scientific reasoning as one of the themes incorporated in science. According to Kind (2013), however, such projects are criticized for giving major emphasis for the domain-general aspect of scientific reasoning (for example TIMSS), lack of categorizing the knowledge and reasoning dimensions explicitly (for example NAEP), lack of explanation of the link between knowledge dimensions and scientific reasoning (for example PISA) and all the three struggle to set an appropriate conceptualization for the science learning domain (particularly for scientific reasoning). Yet, the outcomes of these assessments indicate the effectiveness of a country's educational system and its education quality. Also, the scientific reasoning items are context-dependent and, in some cases, culturally situated to the setting of developed nations. The tests are also mainly meant to assess students' ability in high-income nations, limited to a specific selected common core of science knowledge and skills in keeping the basic interest of the participating countries (Kambeyo, 2017).

The CTSR is one of the commonly administered and used tests by science education researchers (Bao et al., 2009; Lawson, 2004). The test assesses various aspects of scientific and mathematical reasoning; hence, CTSR measures the general attributes of scientific and mathematical reasoning. Osborne (2013) argued due to the domain-general aspect of CTSR, it lacks the considerations of contextual factors. Anderman, Sinatra, and Gray (2012) contended that such tests are limited in measuring students' reasoning abilities and knowledge effectively by considering specific learning situations related to a particular concept in depth. It was also argued that learning is more of domain-specific (science-as-practice) than domain-general (science-as-logic) and developing knowledge is progressive (Gelman & Noles, 2011; Lehrer & Schauble, 2006). Duschl and Grandy (2011) further contended that students' content knowledge learning, ways of reasoning, ways of communicating scientific ideas, and critiques linked to the domain within which learning is taking place.

In addition to this, 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, Mullis, Foy, & Hooper, 2016; Sabanathan, Wills, & Gladstone, 2015). It means the average student in low-income nations is four to five years of learning behind a similar student in the high-income nations. The international assessments allow comparison among participating countries, give opportunities to share techniques, and help to check the educational structures and policy in line with the students' achievement success (Cresswell, Schwantner, & Waters, 2015). Nonetheless, it is difficult to represent every aspect of students' ability by considering all the varying factors such as quality of education, learning strategies, assessment system, class-size, health status, nutrition, and family background. It is difficult to develop a common assessment procedure that suits all countries because there are curricular variations, achievement gaps, and weights given for a particular skill in terms of various interests of a specific country (Au, 2007; Greaney & Kellaghan, 2007). 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, 2010).

These indicate the need to develop items that assess students' reasoning ability by considering the context of the students in terms of the curricula and their learning. Besides, there are low conceptual and reasoning abilities of the Ethiopian upper primary school (grade 7 and 8) students' as reported by the results of the research project "transforming the pedagogy of STEM subjects (TPSS)" (Alemu, Kind, Tadesse, Atnafu, & Michael, 2017) that inspired the researchers to deal with assessing reasoning ability.

The aforementioned problems drove the need for context-based, construct-based, and domain-specific assessment tools that can measure current students' reasoning abilities in low-income nations such as countries in sub-Saharan Africa. Therefore, this study assessed students' reasoning levels along with developing and validating a scientific reasoning assessment tool (SRPT hereafter) for middle school (Grade 8) students.

Hence the study was engaged in answering the following research questions.

(1) To what extent is the SRPT valid and reliable to measure Grade 8 students' scientific reasoning progress?
(2) How do the Grade 8 students' progression trends relate to the different reasoning levels measured by the SRPT?

THEORETICAL FRAMEWORK

The development of an assessment tool to measure scientific reasoning of this kind seeks a guiding framework. As the key concepts are scientific reasoning, and development of the tool, two fundamental frameworks were utilized. These were: 1) the styles of scientific reasoning (SSR) framework represented by four proficiency levels: Generation of claim, Explanations of claim, Evidence-based reasoning and

Table 1. Joint consideration of the SSR and Senocak's frameworks

		Styles of Scientific Reasoning		
		Mathematical deductive reasoning	Model-based reasoning	Experimental reasoning
Levels of Proficiency	Drawing conclusion Evidence-based reasoning Explanations of claim Generation of claim	Items developed for each SSR and proficiency level following Senocak's Framework of: • item formation, • content validation, • construct validation, and • reliability calculation		

Drawing conclusion in their increasing order, and developed from joint consideration of Ford and Wargo (2012) and McNeill and Krajcik (2011) theoretical frameworks, and 2) the development framework of Senocak's (2009) that considers the development of items for each style and level. The Styles of Scientific Reasoning (SSR) framework encompasses six styles of scientific reasoning (SSR) identified through analysis of the history of science: mathematical deduction, experimental evaluation, hypothetical modeling, categorization and classification, probabilistic reasoning, and history-based evolutionary reasoning. But, the first three styles are commonly represented and dominated in primary school science (Hacking, 2012), and Ethiopia follows early compartmentalized delivery of sciences starting from grade 7 that seeks reasoning as a foundation. Therefore these three styles are used to guide the study each of which are leveled into four proficiency levels that are used to identify and demarcate the scientific reasoning construct.

Senocak's framework is a four stages development procedure of a tool. It stages item formation, content validation, construct validation, and reliability calculation. Senocak's framework was used in this study because it allows a construct-driven assessment development procedure by first considering the conceptualization of a construct to be studied.

This study followed the Senocak's framework for the development of SRPT by considering one more stage, pilot testing, to check the consistency between items and the levels, to determine the duration of time required for the test and to make necessary revision. Table 1 presents the matrix of the approach to the study, the details of which are provided in subsequent sections.

METHODS

The method employed for this study included the joint consideration of the four stages of Senocak's Framework with that of the levels of proficiency. The detail of each stage is presented below. To have at least the minimum requirement of the recommended sample size for one parameter Item response theory (IRT) model, which is 200 (Wright & Stone, 1979), six schools were selected randomly for this study. One physics teacher from each school took part in the development of the items. Two hundred and forty-two grade eight students were study targets from the total students of about one thousand and two hundred students. The selection of 8th grade students is because grade 7 is an early stage of the partition of science into three independent subjects and the beginning of physics to be learned as a subject in the Ethiopian curriculum.

Stage 1: Item Formulation

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

This study approaches the construct scientific reasoning from a two-dimensional structure: 1) Styles of Scientific Reasoning (SSR) and 2) proficiency levels. The 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, for example, may require mathematical deductive reasoning with mathematical relationships, model-based reasoning with physics ideas, and data-based reasoning with data. Hence, this study framed the scientific reasoning construct on this three SSR and developed the SRPT that assess grade 8 students' scientific reasoning ability.

Proficiency levels require identifying the progress of students' understanding and reasoning from tacit towards a higher form of understanding. Proficiency levels for this study are identified based on the Ford and Wargo (2012) and McNeill and Krajcik (2011) theoretical frameworks. Ford and Wargo (2012) presented a framework with five sub-categories: Nonact - Recounting - Applying (explain) - Juxtaposing - Evaluating. McNeill and Krajcik (2011) provided a framework consists of four components: claim- evidence - reasoning - rebuttal depending on the experience, understanding, and age of students. The framework of Ford and Wargo (2012) is developed to address conceptual understanding and epistemic knowledge from recount to evaluation level whereas the framework of McNeill and Krajcik's (2011) was developed by incorporating the structure of reasoning.

The framework of Ford and Wargo (2012) gives more focus on the content and epistemic knowledge. They developed a scaffolding framework that shows students

Table 2. Progress of Reasoning in terms of Styles of Scientific Reasoning and proficiency levels

		Styles of Scientific Reasoning			
		Mathematical deductive reasoning	Model-based reasoning	Data-based reasoning	Items
Levels of Proficiency	Drawing conclusion	Draw mathematical model based on the relationship between variables. Scientific knowledge can temporarily be concluded based on the available evidence.	Draw conclusion using Knowledge of scientific concepts, laws and theories. Scientific knowledge can temporarily be concluded based on the available evidence.	Provide inferences based on the relationship between data. Scientific knowledge can temporarily be concluded based on the available evidence.	1.4, 2.4, 3.4, 4.4, 5.4, 6.4, 7.4, 8.4, 9.4, 10.4, 11.4
	Evidence-based reasoning	Generate evidence based on the relationship between variables. A scientific knowledge needs to be supported with various evidence before accepting or rejecting.	Generate scientific evidence to support explanations. A scientific knowledge needs to be supported with various evidence before accepting or rejecting.	Generate reasoning about relationship between data. A scientific knowledge needs to be supported with various evidence before accepting or rejecting.	1.3, 2.3, 3.3, 4.3, 5.3, 6.3, 7.3, 8.3, 9.3, 10.3, 11.3
	Explanations of claim	Use mathematical equations to solve physics problems. A claim needs to be explained based on the relationship between variables.	Identifying relationship between concepts to provide scientific explanations. A claim needs to be explained based on the relationship between variables.	Interpret and give meaning for measurement and observation based on the given data. A claim needs to be explained based on the relationship between variables.	1.2, 2.2, 3.2, 4.2, 5.2, 6.2, 7.2, 8.2, 9.2, 10.2, 11.2
	Generation of claim	Identify given variable, unknown variables and formulas	Remember, state, and recall concepts, theories, laws	Read data from tables and graphs; identify units of measurements	1.1, 2.1, 3.1, 4.1, 5.1, 6.1, 7.1, 8.1, 9.1, 10.1, 11.1

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, the framework of Krajicik and McNeil (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 framework of Ford and Wargo (2012) informs about students' progress of conceptual understanding but it gives less emphasis to the students' pattern of reasoning. The framework of McNeill and Krajcik (2011) considers the structure of reasoning which also gives less emphasis to conceptual understanding, especially content knowledge. Therefore, the combination of both frameworks enabled our development of proficiency levels by incorporating both students' reasoning patterns and three types of domain-specific knowledge.

Based on the amalgamation of these two frameworks, the proficiency levels that depict students' progress of scientific reasoning were developed. The first level of McNeill and Krajcik (2011), claim, can be categorized under the first and the second levels of Ford and Wargo (2012) because both recount and explain are about students' claims in different levels of understanding. Therefore, for this study, the first level is the generation of claim which demands students' describing knowledge based on simple facts and the second level is an explanation of claim by relating and integrating factual knowledge. The third level is evidence-based reasoning based on the McNeill and Krajcik (2011) framework which requests students to generate evidence for the explanations they provided in the second level; and the evidence is based on the juxtaposed knowledge according to Ford and Wargo (2012). The fourth level is concluding by considering a high level of understanding according to Ford and Wargo (2012) and by evaluating available evidence according to the framework of McNeill and Krajcik (2011).

Item development and modification

Items were developed from the contents of grade 7 and 8 physics subject from which students learned to represent the levels in Table 2. Initially, the researchers developed 45 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 contents students have already learned in grade 7 and 8. Items that match the curriculum framework of Ethiopia were selected. The item development followed Haladyna, Downing, and Rodriguez (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.

Table 3. A Sample Item

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

The items developed were given to six selected physics teachers who have been teaching grade eight physics. The teachers provided comments on the setting of the items and if they represent the levels indicated in Table 2. There were also discussions with the teachers about the anticipated students' levels of reasoning and the suitability of the items. Based on the discussions and teachers' comments, the progress levels and the items were improvised, and some new items incorporated. Accordingly, twenty-two items were revised, three items rejected, and five new items included. Finally, 44 items were made available. The final 44 items are developed to represent the levels identified in the table (2) in 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).

A sample item is provided below. It is supposed to measure the data-based reasoning of students. Item (8.1) is designed to assess the factual understanding as a first level, the second level (8.2) requires students to explain the relationship between variables, and in the third level (8.3) they are required to provide evidence as to why they explained the second level by using the data. Finally, in the fourth level (8.4) students are expected to draw a conclusion based on the pattern of the data given in the table

Question8: Two students performed an experiment. To accomplish this, they collected 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 in Table 3.

8.1 Which of the following measurement is **not** value of current measured by the first student?

a) 1.40 A b) 0.73 A c) 2.20 A **d) 1.45 A**

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

a) They want to know the relationship between voltage and current

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

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.

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 less than 2.95 A

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

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

Stage 2: Content Validation

Beyond the participation of school teachers, the following involved ensuring content validity. Two physics instructors, working at Hosanna College of Teacher Education with experience of teaching physics at middle schools, high schools, and college levels; three Ph.D. candidates with experience of teaching physics, specializing in physics education, and who took courses on assessment and scientific reasoning. 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 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. The experts' comments and suggestions for improvement were duly considered when modifying the items during the second round revision. Through this process, 44 final items were made ready for pilot testing.

Stage 3: Construct Validation

The core task of this research is constructing and validating 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.

The items were pilot tested at Hadiyya Zone of Southern Nation, Nationalities, and Peoples Region (SNNPR) in Ethiopia. Fifty students from grade eight participated in the pilot study. This is a sufficient sample size to see what is happening (Wright & Tennant, 1996). Based on initial piloting with those 50 students - who did not involve in the final data collection improvements were made on the items.

After validating the items based on quality indicators supported by expert judgments and the psychometric analysis results from the pilot testing, final data were collected from 242 students. The following discusses the observed results based on the data from the 242 students who took the test. This enhances further the validity of the items developed. 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.

Item separation and reliability of the items, for the pilot testing, were 2.12 and 0.82 respectively. The range of Infit and Outfit values were in the range between 1.05 and 0.9, and 1.24 and 0.89 respectively, which is in the acceptable range. However, items 3.2, 5.4, 3.3, 2.4, 5.1 and 5.1 were items with negative correlation. Items 8.4, 4.2, 2.1, 11.3, 6.3, 11.4, 6.4, 1.3, 5.3, 3.4, 9.4, 5.2, 6.2, 2.3, 1.2, 1.1, 4.3, 4.1 and 11.1 have very low correlation, less than or equal to 0.2. There was also a big gap (about 1STD) between students' mean value and items mean value, which suggests that the items were difficult for the students. There was a big gap between items 8.1 and 7.1 and also between 7.1 and 6.1. The results clearly tell that the test needs some modifications. Based on this result some modifications were made before administrating the final test. The modification was in terms of conceptual arrangement, distracter and grammatical arrangement level. Items 1.1, 3.1, 3.2 and 5.1 were re-worded in suitable grammatical arrangement for the students. Items 4.1, 8.1, 8.2 and 11.1 were modified because they were found to be unclear. The answer options of items 2.3, 2.4, 3.3, 3.4, 5.3, 5.4, 6.2, 6.3, 6.4, 10.3, and 10.4 were revised because they were not answered by the students as expected theoretically.

For the analysis of the items, Rasch model was further used because Rasch analysis allows construct-driven assessment procedures that help to develop progressive levels of a construct (Wilson, 2004). Rasch modeling techniques allows establishing the construct validity by evaluating the fit of the instruments' items to the underlying construct (Bond & Fox, 2007; Martin & Jamieson-Proctor, 2019). In order to ensure the construct validity fit statics, dimensionality and item person map were used.

Fit statistics

Fit statistics provides the discrepancies between the expected Rasch modeling and actual data of the test. To determine the degree to which the data fits the Rasch model, item Infit and Outfit values were determined. Infit detects unexpected patterns of responses on items whereas outfit detects unexpected responses on items which are very easy or very difficult. Mean square residuals (MNSQ) and standardized z-statistics (ZSTD) are ways to provide the results of Infit and Outfit values. The recommended range of infit and outfit of MNSQ for such a kind of multiple choice tests is 0.7 to 1.3 (Bond & Fox, 2007). The acceptable range of ZSTD value is between -2 and 2; however its value is important only when the MNSQ values are not in acceptable range (Linacre, 2016). The values greater than the recommended interval indicates that data are less predictable with respect to the model and values less than the recommended interval indicates that data are more predictable with respect to the model.

For the SRPT, four items (item 7.1, 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 reasoning. The ZSTD values for the items 7.2, 7.3 and 7.4 were 4.6, 3.7, and 2.9 respectively which indicates that the items caused a threat for the validity of the test. The correlations of the items were also poor (less than 0.2). For this reason, the items were rejected from the tool. After rejecting the four items, the Infit MNSQ values fell in the range of .81 to 1.10 which indicates that the learners' abilities match the item difficulties. The Outfit MNSQ values fell in the range between .78 and 1.15 which also indicates that the responses have expected patterns. The ZTSD values for the items fell in the accepted interval (-2 to +2) except for items 3.2 and 4.3.

The fit statistics tells that item 4.3 has an infit MNSQ of 0.85, outfit MNSQ of .83, infit ZSTD of -3.3, and oufit ZSTD of -2.8. Item 3.2 has an infit MNSQ of 0.81, outfit MNSQ of .74, infit ZSTD of -3.6, and oufit ZSTD of -3.4. Hence, the items are over fitting; however, they do not 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 (D. Martin & Jamieson-Proctor, 2019).

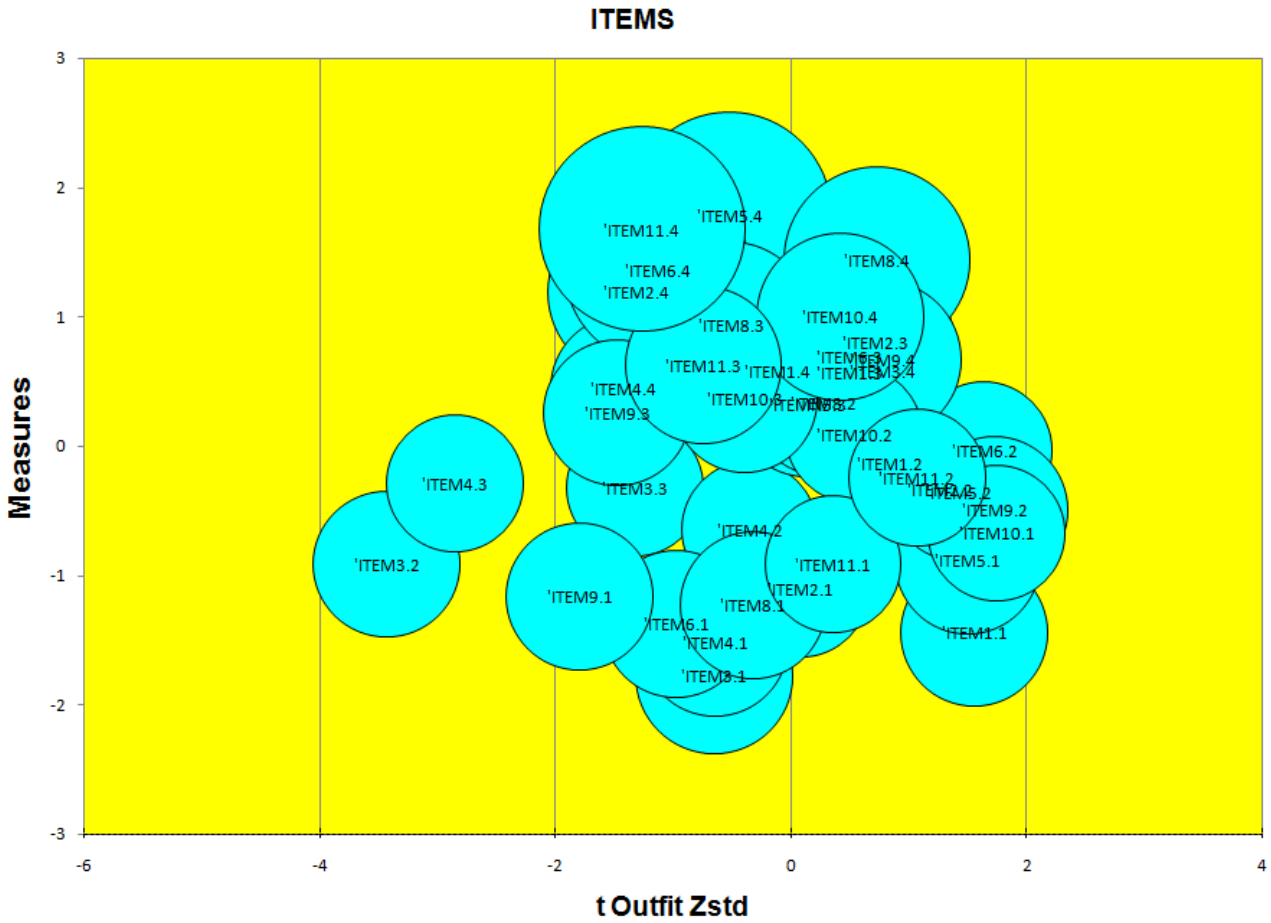

Figure 1. Bubble chart of the SRPT shows how each of the 40 items fit the scale

The items that lie in the left side of the bubble chart show the data is more predictable than Rasch model expectations. The data does not contradict the model rather it indicates that there is redundancy of students' responses. Validity of a scale becomes a threat when there are more misfit (underfitting) items in the right hand side of the bubble chart than items (overfitting) in the left hand side (Martin & Jamieson-Proctor, 2019). The bubble chart in Figure 1 and the values in Table 2 depict that the only items lie outside the Rasch model expectations are item 3.2 and item 4.3. The two items are over fitting because they are placed in the left side of bubble chart. The two items do not affect construct validity. They only indicate that there is redundancy of responses for the items.

The Infit and outfit MNSQ overall values for 40 items were 1.00 and .98 respectively and indicate that the expected mean square value of 1.00 was achieved (Bond & Fox, 2007). The mean of Infit and outfit ZSTD overall values for 40 items were 0 and -.1 respectively which also indicate that the expected ZSTD mean of 0 was almost achieved (see Figure 3). Consequently, the result indicate that the data fit the model and this in turn implies that the overall fit of the SRPT reasonably represent the construct of scientific reasoning and all the items contribute to a single underlying construct.

Dimensionality (PCA on residuals)

The assumption of unidimensionality is one of the most important aspects of Rasch analysis. 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 (2000) 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 and unexplained variance in 1st contrast is recommended not to be greater than 15%. With the help of Principal Component Analysis (PCA), variance was examined in order to identify contrasts in the residuals. This enables to determine whether the data has a secondary dimension or not. A data which has secondary dimension contains items which measure constructs that was not planned to measure.

In the present data of SRPT, the raw variance explained by the items was 23.7% and the first contrast was 5.3% which can be considered as an evidence for unideminsionality. The ratio of the variance explained by measurement to the first contrast is about 5:1, which is also corroboration to the unidimensionality for the SRPT. The result suggests that the test has no significant secondary dimension.

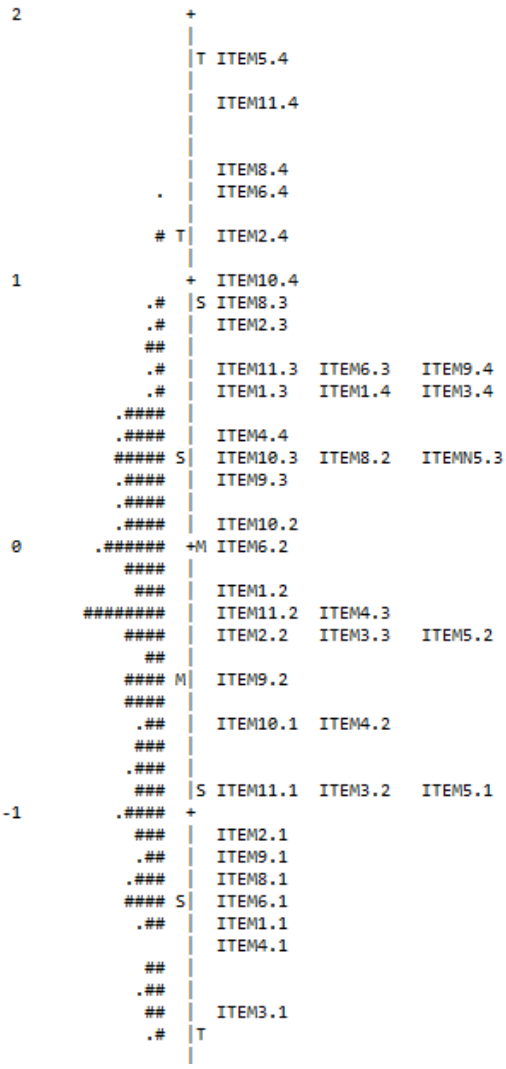

Figure 2. Person-item map for the SRPT: Logit scale with increasing values from bottom to top distributes the person abilities (on the left) and the item difficulties (on the right)

Item person map

Item-person map in Figure 2 represents item difficulty and person ability on the same scale. In Figure 2, the left side shows students and the right side shows items. Each "x" in 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. The Item-person map represents item difficulty and person ability on the same scale called logit scale. The logit scale is a common measurement unit to locate person ability and item difficulty according to person ability and item difficulty estimates (Bond & Fox, 2015). A student 1 logit above or below an item indicates that the person has a chance of 75% answering the item right or wrong respectively. A student placed opposite to the item on the map means that 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 item measure values on the item person map demonstrate that SRPT for all 40 items are ranged from highest measure (item 5.4 with measure of 1.79 logits) to lowest measure (item 3.1 with measure of −1.77 logits). From the 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. Items 8.4, 6.4, 2.4, 10.4, 2.3 and 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 6.1, 4.1, 1.11, 9.1, 8.1, 3.2, 11.1 and 4.2 are ordered in the bottom part of the item person map which is according to theoretical prediction except for item 3.2 and 4.2. In the same way items 11.4, 5.4, 8.4, 6.4, 2.4, 10.4 2.3 and 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 predication in Table 1. 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 level of drawing conclusion. Finally, items 10.2, 10.3, 6.2, 9.3, 1.2, 5.2, 11.2, 2.2, 3.3, 9.2, 10.1, 4.2, 4.3, and 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 higher level of reasoning such as evidence generation and drawing conclusion were answered by 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.

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 the 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 logit. The distance between the items in logit value implies that there is no significant gap among the items. This suggests that students' ability was well targeted with the scale.

The item measures for SRPT were expressed using 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 logit scale that ranged from −2.77 to 1.3. The range of person ability is wider in 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 position of students mean (M) and item mean (M), that

SUMMARY OF 242 MEASURED PERSONS

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	15.6	38.1	-.49	.38	1.00	.0	.98	.0
S.D.	5.9	2.9	.85	.04	.17	1.2	.23	1.1
MAX.	30.0	40.0	1.30	.62	1.46	2.8	1.53	2.7
MIN.	3.0	20.0	-2.77	.35	.55	-3.5	.45	-3.2

REAL RMSE	.40	ADJ.SD	.75	SEPARATION	1.88	PERSON RELIABILITY .78
MODEL RMSE	.39	ADJ.SD	.75	SEPARATION	1.95	PERSON RELIABILITY .79
S.E. OF PERSON MEAN = .05						

VALID RESPONSES: 95.1%

PERSON RAW SCORE-TO-MEASURE CORRELATION = .98 (approximate due to missing data)

CRONBACH ALPHA (KR-20) PERSON RAW SCORE RELIABILITY = .78 (approximate due to missing data)

SUMMARY OF 40 MEASURED ITEMS

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	94.3	230.2	.00	.16	1.00	.0	.98	-.1
S.D.	42.2	6.9	.93	.02	.07	1.2	.11	1.2
MAX.	178.0	240.0	1.79	.21	1.10	1.9	1.15	1.8
MIN.	26.0	209.0	-1.77	.14	.81	-3.6	.73	-3.4

REAL RMSE	.16	ADJ.SD	.91	SEPARATION	5.75	ITEM RELIABILITY .97
MODEL RMSE	.16	ADJ.SD	.91	SEPARATION	5.83	ITEM RELIABILITY .97
S.E. OF ITEM MEAN = .15						

Figure 3. Summary of Person and item Statistics: Reliability Coefficients and Separation Indices

the test was difficult for the sample and it is required more easy items to represent the item difficulty and person ability in a better proportion.

The result from item person map, generally, indicates that the scientific reasoning test seems to fairly measure students' reasoning ability. The 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.

Stage 4: Reliability

The separation and reliability index tells the reproducibility of the items and orderings of persons. Item separation is used to determine how well the items are ordered to separate students in their ability levels. Item separation (> 3) implies items hierarchy is good enough to verify the items difficulty into high, medium and low; item reliability (> .9) implies person sample is large enough to distinguish high, medium and low achievers (Linacre, 2012).

The summary statistics of 40 items and 242 students are shown in Figure 3. 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, Staver, & Yale, 2014). The item reliability (.97) suggests that the SRPT is replicable. That means if one gives the test to another similar group; it has possibility to 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., 2014). 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 a desired continuum based on the difficulty levels.

DISCUSSIONS & CONCLUSIONS

This study was engaged in the development and validation of scientific reasoning progress tool for grade 8 students based on the Senocak's framework. It was found that the SRPT is offered as a valid and reliable instrument that is construct-driven, domain-specific, and higher order thinking assessment that can measure students' reasoning progress. It was found that the SRPT was appropriate and fairly measures students' scientific reasoning progress. SRPT can separate students with low reasoning ability (generation of claim) from high reasoning ability (drawing conclusion) and classify students into low able to high able.

Factual knowledge should not be neglected in science learning since it is the foundation for the construction of

other levels of the reasoning progresses. Students are required to grasp the basics before establishing conceptual understanding. In literature this level is considered as inevitable part of school science for students' healthy and proper progress towards higher level like evidence generation and evaluation (Anderson, Sinatra & Gray, 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.

Level of explanation is characterized by students' attempting to relate particular facts and concepts in order to arrive on general conception of how those facts and concepts give meanings (Corcoran, Mosher & Rogat, 2009). Ford and Wargo (2012) put this level as a level of explanation and characterized by how students describe scientific knowledge to provide explanation about natural phenomena. It was found that the items which were designed to measure the level of explanation are also ordered as expected and established in level of reasoning progress in Table 2. At this stage the items were developed to measure students' abilities of relating concepts or variables to provide meaningful explanations. The tool properly measures students' ability. Most students were able to provide correct explanations for the factual responses. Students were also able to solve physics problems using formula, 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 critical stage for students' real progress towards higher 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, 2011), it was demanding for most grade 8 students to support the claim with valid scientific evidence. Only few students were able to generate scientific evidence for the explanation or claims. The reason why students faced difficulty of generating valid evidence might be related to which evidence require one's creative abilities to match theoretical relationship between concepts, laws, and theories with available data (Kuhn, 2002; Zimmerman, 2000). To provide evidence, students are 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 in helping students to pass the regional exams more than providing activities which invite students to be engaged in evidence-based activities, which is in agreement with previous studies (Anderman et al., 2012; Chinn & Malhota, 2002; Joshi & Verspoor, 2013). The type of assessment provided by teachers also contributes for students' poor ability in generating evidence. The assessments developed by teachers mostly are meant to measure factual and procedural knowledge it lacks assessing students' ability of generating evidence, which is also in line with previous findings (Anderman et al., 2012; Joshi & Verspoor, 2013; Teshome, 2017).

The items that were meant to measure students' ability of *drawing conclusion* were also ordered at the top of item-person map indicating that most students were not able to draw scientific conclusion. It was found that only few students were able to draw conclusion scientifically. The reason for students' low ability to draw conclusion could be related to lack of conceptual understanding and inability to generate valid evidence. In order to draw conclusion students need the ability to compare and contrast evidences from various sources. Why students fail to draw valid conclusion could also be due to lack of practices of such skills in regular classroom activities, and skills of drawing conclusions requires high cognitive abilities (Erlina, Susantini & Wasis, 2018), and requires continuous training and teachers' skills (Hans, 2013). Children face difficulty in generating evidence and drawing conclusion scientifically 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).

Literatures on students' understandings of the nature of science characterize three levels of epistemic development (Cary & Smith, 1993; Ford & Wargo, 2012; Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002). At the lowest level, students are able to carry out scientific reasoning by remembering, describing, and solving conceptual problems. At this level students' reasoning ability is characterized withholding positivist view that distinguishes little between describing and explaining nature. At an intermediate level, students are able to explain their reasoning and provide evidences for the explanations. At this level students tend to express the possibility of multiple explanations for science phenomena, but they may take a view of there exists a single reality. At the most advanced level, students are able to critique scientific reasoning and draw scientific conclusion based on the available evidences. 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 lack of high level reasoning is related with 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, 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 text book and what is told by teachers, and they are unable to relate scientific knowledge with day to day practices (Edmondson & Novak, 1993; Özdemir, 2007).

RECOMMENDATIONS

This article followed a construct-driven assessment model to develop scientific reasoning tool. It was observed in this study that the Rasch model is a simple and effective tool for the analysis of 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 low income countries.

From Figures 1 and 3 one can see that grade 8 students' scientific reasoning ability is low and only a few students' achieved the higher levels of reasoning. Even if the students' reasoning ability was limited largely to the lower levels, there is an indication from the result that students' reasoning ability can be improved towards higher order reasoning if suitable instructions and assessment techniques are preferred that can promote such abilities. Studies reported that children face difficulty in generating evidences 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). Thus, curriculum needs to give emphasis towards inquiry based science education with explicit instruction in nature of science (NOS) as an instructional approach (Duschl & Grandy, 2011; Meyer & Crawford, 2011), dialogical teaching (Osborne, 2010; Osborne, Erduran, & Simon, 2004) along with the proper assessment strategies in order to enhance students' higher reasoning abilities, and contents and materials need to be contextualized. It is recommendable that teacher training institutions and curriculum designers need to incorporate the nature of science in school curriculum and classroom instructions explicitly in order to enhance teachers' and students' view about the nature of science. Students should be informed the modern view of nature of science which emphasizes that science knowledge is tentative; empirical, derived from human inference, imagination, and creativity; socially and culturally embedded (Lederman, 2006). Explicit instruction of nature of science in classroom instruction enables students to think that scientific knowledge is subjective, tentative, socially constructed, cultural 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). In order to enhance students' reasoning it is also required to shift current assessment system which heavily focuses on content based and low ability to incorporate assessments that measure higher reasoning abilities and helpful for scaffolding.

This study focused only on the development and validation of scientific reasoning progress tools that can measure a limited aspect of students' ability of scientific reasoning in physics. The study provides evidence which implies that the middle school students reasoning progress from generation of claim to drawing conclusion. However, a single study with few schools and a single construct cannot clearly provide the whole picture about students' status and levels of reasoning. It is recommended to explore students' pattern of reasoning progress in a wider scope including a qualitative approach to obtain detailed information about students reasoning abilities which can lead to appropriate improvements in the current learning methods and assessment strategies.

REFERENCES

Alemu, M., Kind, P. Tadesse, M., Atnafu, M., & Michael, K. (2017). Challenges of science teacher education in lowIncome nations - The case of Ethiopia. ESERA-17 conference proceedings, Dublin, Ireland.

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. https://doi.org/10.1037/a0021017

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. https://doi.org/10.1080/03057267.2012.655038

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational researcher, 36*(5), 258-267. https://doi.org/10.3102/0013189X07306523

Bao, L., Cai, T., Koenig, K., Fang, K., Han, J., Wang, J., . . . Luo, Y. (2009). Learning and scientific reasoning. *Science, 323*(5914), 586-587. https://doi.org/10.1126/science.1167740

Bond, T. G., & Fox, C. M. (2007). *Applying The Rasch Model: Fundamental Measurement in the Human Science* (2nd Ed). New Jersey: Lawrence Erlbaum.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Dordrecht, Netherlands: Springer. https://doi.org/10.1007/978-94-007-6857-4

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. https://doi.org/10.1080/15248372.2011.567201

Carey, S., & Smith, C. (1993). On understanding the nature of scientific knowledge. *Educational Psychologist, 28*(3), 235-251. https://doi.org/10.1207/s15326985ep2803_4

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education, 86*(2), 175-218. https://doi.org/10.1002/sce.10001

Corcoran, T. B., Mosher, F. A., Rogat, A. D. (2009). Learning progressions in science: An evidence-based approach to reform. Philadelphia, PA: Consortium for Policy Research in Education. Retrieved from http://repository.upenn.edu/cpre_researchreports/53

Davies, M. (2013). Critical thinking and the disciplines reconsidered. *Higher Education Research & Development, 32*(4), 529-544. https://doi.org/10.1080/07294360.2012.697878

DeBoer, G. E. (2011). The globalization of science education. *Journal of Research in Science Teaching, 48*(6), 567-591. https://doi.org/10.1002/tea.20421

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. https://doi.org/10.7771/1541-5015.1538

Duschl, R. A., & Grandy, R. E. (2012). Demarcation in science education: Toward an enhanced view of scientific method. In *Epistemology and Science Education: Understanding the Evolution vs. Intelligent Design Controversy*. Taylor and Francis. https://doi.org/10.4324/9780203839638

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. https://doi.org/10.1002/tea.3660300604

Embretson S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Maheah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Erlina, N., Susantini, E., Wasis, W., Wicaksono, I., & 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), 972-985. https://doi.org/10.33225/jbse/18.17.972

Ford, M. J.,& Wargo, B. M. (2012). Dialogic framing of scientific content for conceptual and epistemic understanding. Science Education, 96, 369-391. https://doi.org/10.1002/sce.20482

Gelman, S. A. (2003). *Oxford series in cognitive development. The essential child: Origins of essentialism in everyday thought.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195154061.001.0001

Gelman, S. A., & Noles, N. S. (2011). Domains and naïve theories. *WIREs Cognitive Science, 2*(5), 490-502. https://doi.org/10.1002/wcs.124

Gott, R., Duggan, S., & Roberts, R. (2008). Concepts of evidence. *School of education: University of Durham*. Retrieved from http://www.dur.ac.uk/rosalyn.roberts/Evidence/cofev.htm

Greaney, V., & Kellaghan, T. (2007). *Assessing national achievement levels in education*: The World Bank. https://doi.org/10.1596/978-0-8213-7258-6

Hacking, I. (2012). 'Language, truth and reason'30 years later. *Studies in History and Philosophy of Science Part A, 43*(4), 599-609. https://doi.org/10.1016/j.shpsa.2012.07.00

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. https://doi.org/10.1207/S15324818AME1503_5

Han, J. (2013). *Scientific reasoning: Research, development, and assessment.* (Electronic Thesis or Dissertation). The Ohio State University. Retrieved from https://etd.ohiolink.edu/

Hanushek, E. A., & Woessmann, L. (2015). *The knowledge capital of nations: Education and the economics of growth*: MIT press. https://doi.org/10.1111/1475-4932.12298

Hewitt, P. G. (2006). *Conceptual physics*. 10th ed. San Francisco: Pearson Addison Wesley.

Joshi, R. D., & Verspoor, A. (2013). *Secondary Education in Ethiopia: Supporting Growth and Transformation*. Washington, DC: World Bank. https://doi.org/10.1596/978-0-8213-9727-5

Kambeyo, L. (2017). Scientific Reasoning Skills: A Theoretical Background on Science Education.

NERA Journal, 14, 40-64. Retrieved from http://doktori.bibl.u/

Kind, P. (2013). Establishing A ssessment S cales U sing a N ovel D isciplinary R ationale for S cientific R easoning. *Journal of Research in Science Teaching, 50*(5), 530-560. https://doi.org/10.1002/tea.21086

Kind, P., & Osborne, J. (2017). Styles of scientific reasoning: A cultural rationale for science education? *Science Education, 101*(1), 8-31. https://doi.org/10.1002/sce.21251

Kuhn, D. (2002). *What is scientific thinking, and how does it develop? The Wiley-Blackwell handbook of childhood cognitive development* (p. 497-523). Blackwell Publishing. https://doi.org/10.1002/9780470996652.ch17

Lamb, S., Jackson, J., & Rumberger, R. (2015). ISCY technical paper: Measuring 21st century skills in ISCY. Victoria University, Centre for International Research on Education Systems [report] https://doi.org/10.4226/80/57f2e5c7a9295

Lawson, A. E. (2004). The nature and development of scientific reasoning: A synthetic view. *International Journal of Science and Mathematics Education*, *2*(3), 307-338. Retrieved from https://link.springer.com/article/10.1007/s10763-004-3224-2

Lederman, N. G. (2006). Syntax of nature of science within inquiry and science instruction. In L. B. Flick, & N. G. Lederman (Eds.), *Scientific inquiry and nature of science. Implications for teaching, learning, and teacher education* (pp. 301-318). Dordrecht: Springer.

Lederman, N. G., 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. https://doi.org/410.1002/tea.10034

Lehrer, R., & Schauble, L. (2006). *Cultivating Model-Based Reasoning in Science Education.* In R. K. Sawyer (Ed.), *The Cambridge handbook of: The learning sciences* (p. 371-387). Cambridge University Press.

Linacre, J. (2016). *A user's guide to Winsteps Ministep Rasch-model computer programs 2016*. Chicago: Winsteps. com.

Linacre, J. M. (2012). *Winsteps® Rasch measurement computer program user's guide*. Beaverton.

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. https://doi.org/10.1080/1743727X.2019.1687669

Martin, M., Mullis, I., Foy, P., & Hooper, M. (2016). *TIMSS 2015 International Results in Science*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. In. Retrieved from http://timss2015.org/wp-content/uploads/filebase/full%20pdfs/T15-International-Results-in-Science-Grade-4.pdf

McNeill, K. L., Krajcik, J. S. (2011). *Supporting Grade 5-8 students in constructing explanations in science: The claim, evidence, and reasoning framework for talk and writing*. New York, NY: Pearson.

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. https://doi.org/10.1007/s11422-011-9318-6

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*. University of Chicago Consortium on Chicago School Research.

Niaz, M. (2017). *Evolving nature of objectivity in the history of science and its implications for science education* (Vol. 46). New York, NY: Springer.

OECD (2016), *Education at a Glance 2016: OECD Indicators*. OECD Publishing, Paris. https://doi.org/10.187/eag-2016-en

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

Osborne, J. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *Science, 328*(5977), 463-466. https://doi.org/10.1126/science.1183944

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

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching, 41*, 994-1020. https://doi.org/10.1002/tea.20035

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), 962-981. https://doi.org/10.1002/tea.21460

Özdemir, G. (2007). The effects of the nature of science beliefs on science teaching and learning. *Uludag University, Journal of the Faculty of Education, 20*(2), 355-372. http://hdl.handle.net/11452/11377

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

Statistics, *4*, 207-230. https://doi.org/10.3102/10769986004003207

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?. *Archives of disease in childhood*, *100*(5), 482-488. https://doi.org/10.1136/archdischild-2014-308114

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. https://doi.org/10.1007/S10956-009-9173-3

Teshome, N. B. (2017). Classroom Participation and Development of Student Attitudes: A Study of Active Learning Practices in Ethiopian Primary Education. *International Journal of Humanities Social Sciences and Education (IJHSSE), 4*(3), 67-68. https://doi.org/10.20431/2349-0381.0403008

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. *Int J of Sci and Math Educ, 15*(4), 663-682. https://doi.org/10.1007/s10763-016-9723-0

Verspoor, A. M. (2008). *At the crossroads: choices for secondary education in Sub-Saharan Africa*: World Bank Publications, The World Bank, number 6537, Juni.

Wilson, M. (2004). *Constructing measures: An item response modeling approach*: New York, Routledge. https://doi.org/10.4324/9781410611697

Wright, B. D., & Stone, M. H. (1979). Best Test Design: Rasch Measurement. Chicago, IL: Mesa Press.

Wright, B., & Tennant, A. (1996). Sample size again. *Rasch Measurement Transactions, 9*(4), 468.

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. https://doi.org/10.1002/tea.20368

Zhou, S., Han, J., Koenig, K., Raplinger, A., Pi, Y., Li, D., Xiao, H., Fu, Z., & 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. https://doi.org/10.1016/j.tsc.2015.11.004

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

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