JOURNAL OF BALTIC SCIENCE EDUCATION
ISSN 1648-3898 /Print/
ISSN 2538-7138 /Online/

EVALUATING SCIENTIFIC REASONING ABILITY: THE DESIGN AND VALIDATION OF AN ASSESSMENT WITH A FOCUS ON REASONING AND THE USE OF EVIDENCE

Ma Luo,
Zuhao Wang,
Daner Sun, Zhi Hong Wan,
Liying Zhu

Abstract. *Scientific reasoning ability (SRA) is widely recognized as an essential goal for science education. There is much discussion on the design and development of assessment frameworks as viable tools to foster SRA. However, established assessments mostly focus on the level of students reasoning attainment. Student ability to use evidence to support reasoning is not adequately addressed and evaluated. In this study, the 6-level SRA assessment framework was conceptualized and validated iteratively via synthesizing literature and a Delphi study. Guided by the framework, an SRA assessment tool adopting and adapting PISA test items and self-created items was developed and administered to 593 secondary students (including 318 8th Graders and 275 9th Graders) in mainland China. Pearson correlation analysis of SRA assessment score and their scores in scientific reasoning provided criterion-related validation for the former (Pearson correlation = .527). Rasch analysis conducted further confirmed the validity and reliability of the SRA test and the assessment framework. Combing quantitative and qualitative methods, the study provides a valid and reliable analytical framework of SRA. It can inform the design of SRA assessments in various science education contexts for diversified audiences.*

Keywords: *Complexity of Scientific Reasoning, Evidence in Reasoning, Rasch Modeling, Scientific Reasoning Ability (SRA).*

Ma Luo
Ningbo University, China
Zuhao Wang
East China Normal University, China
Daner Sun, Zhi Hong Wan
Education University of Hong Kong, China
Liying Zhu
Zhejiang Normal University, China

Introduction

The study of scientific reasoning ability (SRA) is one of the frequently discussed topics in science education. In decades, researchers and educators have devoted great efforts to exploring how students develop their scientific reasoning ability and how their scientific reasoning ability affects their learning achievements (Coletta, Phillips & Steinert, 2007; Ding, 2018; Johnson & Lawson, 1998). In science education, the development of scientific reasoning ability is an essential goal that has long been pursued and prioritized (Bao et al., 2009; Engelmann et al., 2016). Thus, one of the primary tasks of science education is to cultivate students into good reasoners and become scientific literate (Lawson, 2004). K-12 science educators have also made great efforts to foster students' scientific thinking and reasoning through their engagement in familiar phenomena in daily life contexts (Kind & Osborne, 2017; van der Graaf et al., 2019). To cultivate student reasoning ability in and for science learning, many studies have investigated the nature of reasoning (Driver et al., 1994) and its development via teaching practices (Lawson, 2004; Zimmerman, 2000; 2007). Leveraging these research efforts, a batch of assessments of scientific reasoning using standardized tests have been constructed and implemented. These assessments, which mainly focused on the evaluation of the level and complexity of reasoning involved in the processes of solving science problems, can inform and improve educational practices of science (Kind, 2013; Kalinowski & Willoughby, 2019; Lee & She, 2010). However, the valid and reliable assessments of scientific reasoning ability are still needed to diagnose research gaps and identify areas of improvement in science learning and teaching.

The recent discussion on scientific reasoning in science education uncovers and underscores the significance of evidence use. Evidence is viewed as the premise and basis of valid reasoning in the fields of logic and cognitive psychology (Toulmin, 2003), and the collection and analysis of evidence are necessary and vital for the formation of scientific reasoning and thinking (Kanari & Millar, 2004). Scientific evidence and evidence-based reasoning should be the core of students' science learning experiences (Duschl, 2003).

https://doi.org/10.33225/jbse/20.19.261

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

The framework for K-12 science education adopted in the United States regards "engaging in argument from evidence", which was based on the reasoning ability to evaluate evidence about correlation and cause, as one of the eight major practices in science and engineering. The ability and processes of collecting data based on the observation and formulating evidence are key to scientific inquiry and should be emphasized (National Research Council, 2012).

However, in existing scientific reasoning assessments, how students use evidence to support their reasoning processes is inadequately considered and evaluated (Osborne, 2013). Most of the reasoning assessments were based on the use of argument pattern (Toulmin, 2003), and test with written assignments (Yanto et al., 2019; Zhou et al., 2016), classroom discussion (Osborne et al., 2004), and student interviews (Adey & Csapó, 2012; Jimenez-Aleixandre et al., 2000). These assessments revealed students' ability to make correct claims and rational conclusions but did not reflect their ability to capture evidence (Sandoval & Millwood, 2005). In a series of studies, an analytic framework of Evidence-Based Reasoning (EBR), which also based on Toulmin's argument pattern was established and applied to assess students' ability to reason from evidence using writing tasks and classroom discussion. According to this EBR framework, the reasoning is the processing of two kinds of inputs (i.e. data and premise) through three steps of data analysis, evidence interpretation, and rules application to form a claim as the final output (Brown et al., 2010a). Empirical data proved the validity of the framework and affirmed the possibility to evaluate students' SRA based on evidence (Brown et al., 2010b).

Motivated by both the inadequacy and achievement of previous research efforts, the present study aimed to conceptualize and validate an assessment framework with design and development of an assessment tool of SRA that combines the complexity of reasoning and the use of evidence. Specifically, the SRA framework was developed based on the analysis of existing assessment models and the use of a Delphi study which engaged experts in science education. Building on the SRA framework, an assessment tool that incorporated test items from PISA and self-developed items was compiled and implemented. Pearson correlation analysis of the SRA test scores of 593 secondary students (including 318 8th Graders and 275 9th Graders) and their scores obtained in a classic test of scientific reasoning (i.e., Lawson's Classroom Test) provided criterion-related validation for the SRA assessment. And the validity and reliability of the SRA assessment framework and test were further confirmed by Rasch analysis results. The assessment developed in this study provides a valid analytic framework of SRA that can contribute to the design of assessments and educational practices of SRA across a wide spectrum of educational settings.

Theoretical Framework

Scientific Reasoning: Grounded in Evidence

In the perspectives of science education, guided by the goal of cultivating students to be "better reasoners in a general sense and become scientifically literates" (Lawson, 2004), scientific reasoning refers to the ability to systematically investigate a problem concerning science, formulate hypotheses, and test them, control and manipulate variables, and evaluate experimental outcomes, such as data results, and make explanations (Bao et al., 2009; Zimmerman, 2000; 2007).

Even though researchers held different opinions based on various perspectives, it was believed that scientific reasoning is constrained by laws (Moshman, 1998) that should deliberately consider the contextual correlations among assorted information, and consciously coordinate theory and evidence (Kuhn et al. 1995). From the perspective of scientific discovery, Klahr and Dunbar (1988) proposed a dual search model to describe scientific reasoning as search in hypothesis space and an experiment space. The model includes three components: search hypothesis space; test hypothesis; and evidence evaluation. That was similar to Kuhn's (1995) phases of knowledge acquisition, including evidence accumulation and evaluation to provide an explanation and conclusion.

By strongly supported evidence and premise, students arrive at sub-claims if they are involved in a more complicated reasoning process (van Eemeren et al., 2002). Each sub-claim should be connected with appropriate evidence, be it new or additional. In such a way, students will accomplish reasoning and succeed in argumentation (Belland et al., 2008). In other words, what enlightens us is that conducting SR refers to an integrated process whose complexity depends on specific problems to be solved, and such ability enables an individual to persuade the audience (i.e.teachers, peers, and community leaders) to agree with their claims or solutions to specific problems (Hmelo-Silver, 2004).

https://doi.org/10.33225/jbse/20.19.261

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Conceptualizing the Framework for Assessing SRA: Evidence Complexity and Reasoning Complexity

In science research, evidence is usually derived from observations and experiment results to support, modify, refute, or form scientific hypothesis or theories (National Research Council, 2012), and should be accurate, confirmed, and leading to necessary consequences that are "distinct from observations on which it is based on and the principle it is intended to illustrate" (Brown et al., 2010a). In educational settings where science-related problems are to be solved (e.g. the classroom, laboratory or outside of school), the evidence is used for supporting claims or decision-making. And, as reflected in OECD reports, the contextual information, or the evidence, provided could impact how students solve the scientific problems (2006; 2016), and thus can be used to measure the levels of student SRA. The levels of students' engagement and achievement in SRA are influenced and reflected by the nature and complexity of the problem context as defined by 1) its familiarity (i.e. familiar context vs unfamiliar context) (Choi & Hannafin, 1997), and 2) the explicitness (i.e. explicit evidence vs implicit evidence) (Salgado, 2016), and 3) the quantity of evidence (i.e. single evidence vs multiple evidence) embedded in. When there is only one piece of evidence (i.e. single evidence) that can be directly captured rather than deeply explored (i.e. explicit evidence) from the familiar context that resembles daily life experiences, the reasoning processes involved in the problem-solving processes are at the lowest level as the evidence engaged has least complexity. When students deal with multiple, implicit evidence in an unfamiliar context, they are involved in the most complex scientific reasoning processes (Dolan & Grady, 2010; Zhou et al., 2016).

In cognitive science, many studies have investigated the extent to which they are capable of rational thought or acting rationally in different circumstances (Kyllonen & Christal, 1990). As mentioned before, most of the assessment frameworks of reasoning are based on Toulmin (2003)'s: the use of argument pattern. From the perspective of scientific learning and teaching, the level of reasoning refers to student ability to systematically investigate a problem concerning science, formulate and test hypotheses, control and manipulate variables, evaluate experiment outcomes (e.g. data results), and make explanations (Bao et al., 2009; Zimmerman, 2000; 2007).

Grounded in theories of situated cognition, Dolan and Grady (2010) adopted a case study approach to construct a matrix for evaluating the Complexity of Scientific Reasoning during Inquiry (CSRI) by analyzing teaching practices. In this CSRI matrix, students' cognitive processes of reasoning were categorized into four continuing levels based on complexity: least, somewhat, more, and most complex reasoning. Another significant analytical framework of scientific reasoning, the Lawson's Classroom Test of Scientific Reasoning (LCTSR) has been widely adopted and applied since its conceptualization in 1978 and its further improvement in 2000 (Bao et al., 2009; Lee & She, 2010; Thompson et al., 2017). LCTSR, a paper-pencil based assessment composed of 12 paired, two-tier, multiple-choice test items, investigated SRA from six dimensions, which include conservation reasoning, proportional reasoning, control of variables, probability / probabilistic reasoning, correlation reasoning, and hypothetic-deductive reasoning.

From a logical perspective, based on statements or propositions quantity, two kinds of inferences are distinguished. An immediate inference is an assumption, without intervening or "mediating" premises; a mediate inference is a logical inference drawn from more than one premise (Churchill, 1990). As the valid form of inference is the concerned and primary issue of logic, in this study, we shall pay more attention to the content of premises (evidence) and the relationship between premises (evidence) and conclusion. The content of premises represents the context and evidence sources, and the latter is the basic requirement of the reasoning process. We, therefore, focused our exploratory efforts on the holistic and scientific process of reasoning based on evidence to solve problems, and defined SR process as two aspects: direct reasoning and indirect reasoning, which is different from the definition and clarification in logic. In direct reasoning, the relationship between the evidence presented in the context of the science problem is quite simple and involves less complexity. The evidence involved in direct reasoning can be either single or multiple. The reasoning processes are more complicated if multiple evidence is coordinated. Yet in indirect reasoning, students manage complicated relations among multiple evidence, be it covert or overt (Leron, 1985). Such processes demand greater analytical and integrative skills. Table 1 presents the definitions of 3 levels of reasoning based on its complexity. Levels 1 and 2 indicate the direct reasoning requirement. However, level 2 is based on multiple pieces of evidence. Level 3, an advanced level of reasoning, which deals with multiple pieces of evidence with complicated connections and demands students' analytical and integrative skills.

https://doi.org/10.33225/jbse/20.19.261

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Table 1
The Reasoning Complexity: level and definition

Reasoning complexity level	Definition
Level 1: Direct reasoning - 1	Students recognize and extract single evidence (S) from the context, reasoning directly based on evidence.
Level 2: Direct reasoning - 2	Students recognize and extract multiple pieces of evidence (M) from the context, establish simple relations between/among the evidence.
Level 3: Indirect reasoning	Students recognize and extract multiple pieces of evidence (M) from the context, establish complicated relations between/among evidence.

Evaluating SRA Combing the Complexity of Reasoning and Evidence

Integrating evidence complexity (EC) and reasoning complexity (RC), the framework for assessing and defining student SRA in science learning, particularly in science-related problem solving, was initially constructed based on four indicators, they are 1) context familiarity (familiar context vs unfamiliar context), 2) evidence explicitness (explicit evidence vs implicit evidence), 3) evidence quantity (single evidence vs multiple evidence), and 4) reasoning complexity (direct reasoning vs indirect reasoning) that collectively measure the complexity of scientific reasoning. For details of the initial SRA assessment framework, please refer to Table 2. In Table 2, the EC in eight kinds of combinations is matched to the different levels of RC. The CSR level has been formulated in a connective way. As discussed above, the unfamiliar context will improve the complexity of evidence but not as much as the impact of implicit evidence. Thus, SEU and SEF refer to the same level of CSR (Level 1a). CSR is divided into nine levels by the complexity levels of reasoning.

Table 2
The Complexity of Scientific Reasoning (CSR) framework (initial version)

Reasoning complexity	Evidence complexity	Remarks	Level of CSR
Level 1 (Direct reasoning -1)	SEF: Single-Explicit-Familiar	The lowest level of complexity	Level 1a
	SEU: Single-Explicit- Unfamiliar	Unfamiliar evidence adds a little complexity a little	
	SIF: Single-Implicit-Familiar	Implicit evidence adds to the complexity (more than U)	Level 1b
	SIU: Single-Implicit-Unfamiliar	I&U add to the complexity	
Level 2 (Direct reasoning -2)	MEF: Multiple-Explicit-Familiar	Establish simple relations	Level 2a
	MEU: Multiple-Explicit-Unfamiliar	Establish simple relations; U adds a little complexity	
	MIF: Multiple-Implicit-Familiar	Establish simple relations; I adds to the complexity	Level 2b
	MIU: Multiple-Implicit-Unfamiliar	Establish simple relations; I&U add to the complexity	
Level 3 (Indirect reasoning)	MEF: Multiple-Explicit-Familiar	Establish complicated relations	Level 3a
	MEU: Multiple-Explicit-Unfamiliar	Establish complicated relations; U adds to the complexity	
	MIF: Multiple-Implicit-Familiar	Establish complicated relations; I adds to the complexity	Level 3b
	MIU:Multiple-Implicit-Unfamiliar	Establish complicated relations; I&U add to the complexity	

https://doi.org/10.33225/jbse/20.19.261

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Research Methodology

A Delphi study was conducted to improve the content validity and face validity of the initial CSR framework proposed above (Osborn, 1963). Then an assessment tool using standardized test items was further developed to evaluate students' SRA in solving science-related problems.

The Delphi Method: The Modification of CSR Framework

Following the Delphi method, expert opinions in collective intelligence were consulted to elaborate and improve the initial CSR framework. The first stage involved a brainstorming session that gathered selected experts "face to face" to provide opinions and comments (Isaksen, 1998). The expert group comprised of researchers and teachers specializing in science education. Altogether, one professor, four associate professors, two senior lecturers, one middle school principal, three science teachers, seven doctoral candidates, and several master students were recruited as the experts (22 in total).

After the brainstorming session in which diversified opinions were mined and organized, another expert group was formed for the Delphi survey via e-mail. This group comprised of experts in the field of science education including four associate professors, one senior lecturer, six science teachers, and five doctoral candidates. Following the Delphi principles, all experts were positioned "back to back" to ensure the opinions elicited were independent rather than being influenced by each other (Rowe & Wright, 2001). The Delphi survey was conducted in three rounds. The first was an open consultation during which the selected experts shared their suggestions for and comments on the CSR framework. These responses were summarized and returned to the experts for the second round of clarification and commenting. In the following, the same process was administered, and the agreement was reached among all the experts. The reaching of consensus among the experts marked the achievement of the goal of the Delphi study (Bolger & Wright, 2011).

Based on the insights obtained in the Delphi study, the initial version of CSR framework was revised accordingly (Table 3). According to the experts, during the reasoning processes, implicit evidence would add more complexity than unfamiliar context would; reasoning involving familiar context would be less complicated than with unfamiliar context, but such difference in complexity was not obvious enough to distinguish them into different levels of CSR. Also, for different students, the degree of familiarity with the context of a science problem would be different due to their unique life experiences. The identification of the context of a science problem as familiar or unfamiliar would be difficult. Therefore, context familiarity would not be an appropriate indicator of CSR. According to these opinions, the revised framework categorized students' SRA into six levels.

Table 3
The Complexity of Scientific Reasoning (CSR) framework (revised version)

Reasoning complexity	Evidence complexity	Remarks	Level of CSR
Level 1 (Direct reasoning-1)	SEF: Single-Explicit-Familiar	Unfamiliar evidence adds a little complexity	Level 1a
	SEU: Single-Explicit-Unfamiliar		
	SIF: Single-Implicit-Familiar	Implicit evidence adds to the complexity (more than U)	Level 1b
	SIU: Single-Implicit-Unfamiliar	I&U add to the complexity	
Level 2 (Direct reasoning-2)	MEF: Multiple-Explicit-Familiar	Establish simple relations	Level 2a
	MEU: Multiple-Explicit-Unfamiliar	Establish simple relations; U adds a little complexity	
	MIF: Multiple-Implicit-Familiar	Establish simple relations; I adds to the complexity	Level 2b
	MIU: Multiple-Implicit-Unfamiliar	Establish simple relations; I&U add to the complexity	
Level 3 (Indirect reasoning)	MEF: Multiple-Explicit-Familiar	Establish complicated relations	Level 3a
	MEU: Multiple-Explicit-Unfamiliar	Establish complicated relations; U adds to the complexity	
	MIF: Multiple-Implicit-Familiar	Establish complicated relations; I adds to the complexity	Level 3b
	MIU: Multiple-Implicit-Unfamiliar	Establish complicated relations; I&U add to the complexity	

https://doi.org/10.33225/jbse/20.19.261

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

The Development of the SRA Assessment Tool

For evaluating students' general abilities such as SRA, the test designed and developed should not involve content knowledge as most established tests (e.g., LCTSR) did (Bao et al., 2009). As the test may be administered to students with different grades, the knowledge they acquired would be an extraneous variable and should be controlled accordingly. Thus, in designing the SRA assessment tool, the context of a science problem should not involve knowledge learnt from school but provide ample information for students to analyze, capture and transform evidence based on such information. In other words, the designed test items should involve abundant, varied contextual information that is related to science and represented as different forms of "evidence", but do not contain specific science knowledge. This constitutes the most important principle that guided the test design. Also, the test items should be developed according to the CSR framework and more than one item should be included for each complexity level in case of the presence of inappropriate items.

Besides self-developed items, the SRA test also incorporated PISA questions which evaluate students' general science ability (OECD, 2006; 2015; Tamassia & Schleicher, 2002) and highly emphasize the problem context (Bybee et al., 2009; Fensham, 2009), and have been validated via several empirical studies (e. g, Dohn, 2007; Sadler & Zeidler, 2009). In total, the SRA test consisted of 25 items, with 12 multiple-choice questions and 13 constructed response questions. All the items were reviewed by experts in science education to ensure their content validity. Table 4 provides a summary of the complexity level of scientific reasoning involved in each test item.

Table 4
SRA test items and the corresponding complexity level based on the CSR framework

Level of CSR	Evidence complexity	Items
Level 1a	SEF	P01, P04
	SEU	P08
Level 1b	SIF	P09
	SIU	P07, P20, P12
Level 2a	MEF	P13, P17
	MEU	P18
Level 2b	MIF	P05, P21
	MIU	P02, P06, P11
Level 3a	MEF	P10, P14, P19
	MEU	P22, P24
Level 3b	MIF	P03, P25
	MIU	P15, P16, P23

Procedures of Evaluation Study of SRA test

A small-scale pilot study was carried out to explore inadequacies of the SRA test. Through convenient sampling, 31 students including 16 6th Graders and 15 17th Graders; 12 boys and 19 girls) from different schools in Shanghai, China participated in the pilot test which was administered during an out-of-school activity by a science teacher and assisted by one doctoral candidate and two master students in science education for observation. After the test, there was a semi-structured interview to collect student feedback on the test. The data collected (including student test scores, observations, and student feedback) was used to modify the instrument. The SRA test was validated through being done by a much larger pool of students. The scores students obtained in the SRA test were compared and correlated to their scores obtained in the LCTSR, a widely adopted and extensively corroborated assessment for evaluating scientific reasoning (e.g., Bao et al., 2009; Lee & She, 2010). As SRA is a

https://doi.org/10.33225/jbse/20.19.261

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

construct closely related to scientific reasoning, and LCTSR a valid assessment, if a student scores in the SRA test correlate well with their LCTSR scores, this would defend the criterion-related validity of the SRA assessment.

The participants of the LCTSR and SRA test were Grade 8 and 9 students from a junior middle school in Chizhou, China. Altogether, 582 students (309 8th Graders and 273 9th Graders; 306 boys and 276 girls) and 593 students (318 8th Graders and 275 9th Graders; 306 boys and 287 girls) participated in the LCTSR and SRA test respectively. The two tests were done with a one-week interval. During the two 30-minute tests, each student did the printed test paper independently. Before the tests, all participants and their schools were invited to sign the consent form. After getting the consent approvals of the participants, the relevant information of the test was introduced.

Rasch analysis was employed to examine whether the test really targeted and evaluated the construct of SRA and whether the CSR framework and the SRA test matched and corresponded to each other. Rasch modeling is based on Item Response Theory (IRT), a "psychometric technique [that] was developed to improve the precision with which researchers construct instruments, monitor instrument quality, and compute respondents' performances" (Boone, 2016). In Rasch measurement, raw scores are to be converted into logarithmically scaled measures of interval levels. Estimates of personal ability (i.e., student SRA in the present study) and item difficulty (i.e., the difficulty level of SRA test items in the present study) can thus be placed together on a single continuum. As measures of items and persons are sample and item independent, comparisons can be made regardless of the sample chosen or items selected for assessment as long as they measure the same construct (Bond & Fox, 2007). Due to the simple yet strong rationale, Rasch modeling has been extensively adopted in psychometric research on the development and validation of measurement instruments (Wei et al., 2014). The popularity of Rasch model in science education research provides reference for test development and analysis on teacher classroom performance and student learning achievement (Liu, 2010; Randall, 2010; Wei et al., 2014).

According to the assumptions of Rasch modeling, multiple rounds of the test are required to measure data fitness, and the collected empirical data should meet the specified criteria and structure for objective measurement (Liu, 2010; Linacre, 2006; 2011). In this study, the holistic and iterative development process implemented, including 1) pilot test and test refinement, 2) main study of LCTSR and SRA test; 3) Rasch analysis and further refinement (if any) could provide ample empirical evidence to validate and improve the SRA assessment, as well as exemplify its application in assessing SRA in K-12 education settings.

Research Results

The Results of Pilot Test

Data analysis of student scores in the pilot test using SPSS 22.0 revealed satisfying reliability of the SRA test (Cronbach's α=.706). It was appropriate for students at different grades as their scores were distributed (mostly ranging from 9 and 25; *Mean*=16.41, *SD*=4.492), and they could all finish the test within the designated time. Student feedback reflected the SRA test was for evaluating their general "thinking" or "intelligence", not for specific science knowledge. Students also reflected that some of the test items were difficult to understand, which negatively impacted their score. In the following, the phrasing of these items was revised accordingly.

The Results of the LCTSR and SRA Test

Altogether, the test scores of 552 students were put into analysis. The correlation between their LCTSR scores and SRA test scores was .527 (Pearson coefficient, $p = .000$, N=552). Such correlation indicated that, in accordance with LCTSR, the SRA test possessed good and statistically significant pragmatic validity and would help measure students' SRA. Additionally, the Cronbach's α of the SRA test was .809 (N=593), indicating good reliability of the test. It proved that all the 25 test items were measuring the very same construct of SRA. The highest score obtained was 30 (full score) and the lowest was 1 (*Mean*=15.82, *SD*=5.917). Although girls, in general, scored lower than the boys did (boy: *Mean*=15.90, *SD*=6.042, N=306; girl: *Mean*=15.74, *SD*=5.790, N=287), such gender difference was not significant ($t = .341$, $p = .733$).

https://doi.org/10.33225/jbse/20.19.261

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

The Results of Rasch Analysis

In the following Rasch modeling, the overall analysis, item uni-dimensionality test, item fitness, and item distribution were operated using WINSTEP 3.72.0 software and with reference to the user manual (Linacre, 2006; 2011) and complementary studies (Liu, 2010; Sondergeld & Johnson, 2014) for indices criteria. Table 5 presents a summary of all persons and items analysis results. The estimated ability value of all participants is 0.35 (in logits), a little higher than the total item difficulty, which is generally set as 0. In Rasch modeling, the Error reflects the accuracy of parameter estimation, and the closer the Error to 0 (in logits), the better the test. Based on the suggested criteria, person ability and item difficulty error are both acceptable, yet could be improved upon. Person separation index and reliability coefficient are acceptable as well, indicating good item indices. The MNSQ and ZSTD values are nearly perfect as both INFIT and OUTFIT were between 70 and 1.3. The results confirmed that the SRA test was reliable and accountable for measuring students' SRA.

Table 5
Summary of person and item estimates and indices

	Measure	Error	Infit		Outfit		Separation	Reliability
			MNSQ	ZSTD	MNSQ	ZSTD		
Person	0.35	0.47	1.00	0.0	1.03	0.0	1.92	0.79
Item	0.00	0.10	1.00	0.0	1.02	0.1	9.98	0.99

Note: value all in logits.

Table 6 presents the fitness of all the 25 SRA test items. Overall, these statistics fit the Rasch model, further confirming the validity of the SRA test. The standard error (S.E.) for all items is below 1.0 (in logits), ranging from .07 to .14. The Outfit and Infit index of most items are all acceptable as the MNSQ of all items were below 1.3, only except for P24. For P24, the ZSTDs are both -9.9 and the MNSQs are both below 0.5, suggesting the need for revision or even elimination of the test item. Another important value, PT-MEASURE CORR (i.e., the partial correlation between the scores student obtained on the specific item and their total scores) also helps justify the test items. According to Liu (2010) and Linacre (2011), the more positive the correlation, the better the test instrument design is. For the SRA test, all partial correlations are positive, ranging from 0.24 to 0.60, demonstrating acceptable convergent validity of the SRA test.

Table 6
Item fitness of the SRA test

Item	Measure	Model S.E.	Infit		Outfit		PT-MEASURE CORR.
			MNSQ	ZSTD	MNSQ	ZSTD	
P09	-0.58	.10	1.14	3.0	1.3	4.0	0.25
P11	1.6	.10	1.11	2.1	1.29	3.1	0.25
P12	-1.31	.08	1.22	3.4	1.09	1.2	0.52
P01	-1.1	.11	1.15	2.5	1.21	2.2	0.24
P04	-1.93	.14	0.97	-.3	1.21	1.3	0.33
P22	0.54	.07	1.2	3.5	1.2	3.3	0.46
P10	1.07	.09	1.11	2.8	1.19	2.8	0.29

https://doi.org/10.33225/jbse/20.19.261

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Item	Measure	Model S.E.	Infit		Outfit		PT-MEASURE CORR.
			MNSQ	ZSTD	MNSQ	ZSTD	
P06	0.33	.09	1.01	.4	1.14	2.7	0.38
P15	1.62	.10	0.94	-1.3	1.13	1.4	0.39
P23	-0.19	.13	1.11	2.1	1.09	1.2	0.34
P18	-0.16	.09	1.04	1.2	1.07	1.3	0.37
P19	0.25	.09	1	.1	1.05	1.0	0.4
P21	0.32	.07	1.04	.8	1.01	.1	0.59
P14	0.56	.09	1.02	.5	1.04	.8	0.39
P02	0.48	.09	1.02	.5	0.99	-.1	0.4
P03	0.27	.09	1.01	.4	1.02	.3	0.39
P16	1.91	.11	1.01	.2	1	.1	0.33
P07	-0.18	.09	0.98	-.6	0.92	-1.4	0.44
P08	-1.27	.11	0.95	-.8	0.96	-.3	0.4
P13	-0.38	.07	0.96	-.8	0.92	-1.4	0.6
P05	0.21	.09	0.9	-3.0	0.87	-2.8	0.5
P17	-0.99	.11	0.9	-1.7	0.83	-1.9	0.48
P20	-0.03	.10	0.89	-3.2	0.85	-2.9	0.51
P25	-1.81	.14	0.84	-1.8	0.63	-2.6	0.47
P24	0.74	.07	0.43	-9.9	0.49	-9.9	0.54

Note: value all in logits.

In the principal components analysis (PCA) in Rasch modeling, the first eigenvalue is 1.8 (< 2.0), meaning the SRA items are treated unidimensionally (Linacre, 2011). This result showed the SRA test was only measuring the construct of SRA. Figure 1 presents the item loading scatterplot derived from the PCA. The scatterplot shows the contrasts by plotting the loading on each component against the item calibration (Linacre, 2006). For a test instrument, if the contrast loading of every test item falls within the range of -.4 to +.4, the unidimensionality requirement is satisfied. As figure 1 shows, only three items: A(P12), B(P13), and C(P21), had a contrasting loading that fell out of the range. Overall, the SRA test met the unidimensionality requirement, providing further proof for its construct validity (Liu, 2010).

https://doi.org/10.33225/jbse/20.19.261

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Figure 1
Contrast loadings of residuals (standardized residual contrast plot) in principal components analysis

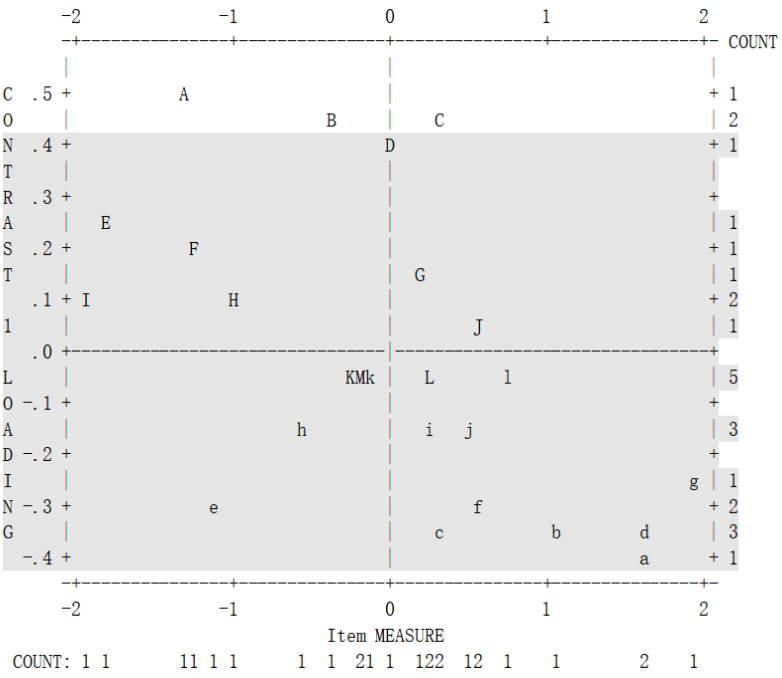
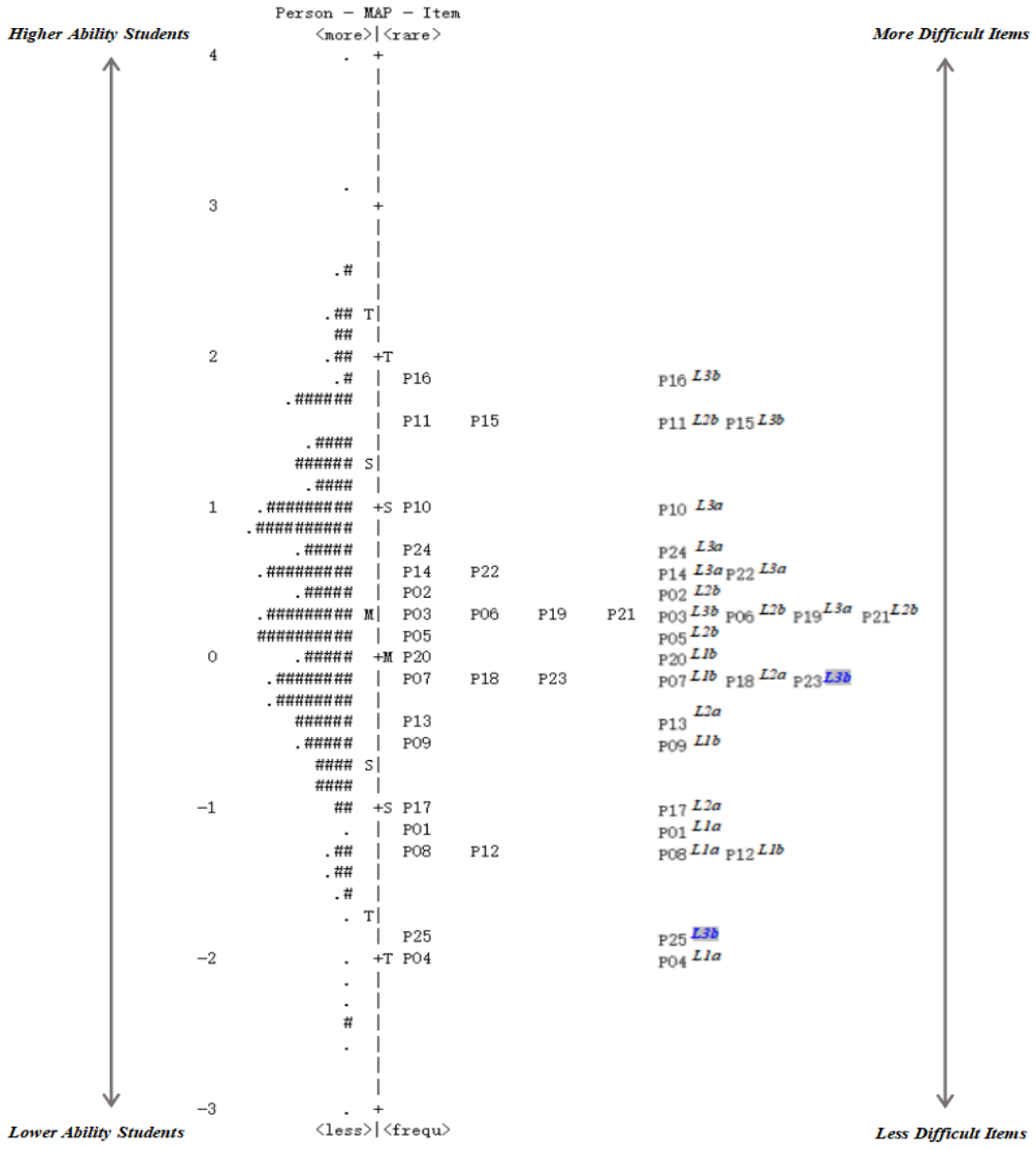


Figure 2 is the combined person and item estimate map, called person-item map or Wright map. It displays the distribution of persons and items on the same interval, logarithmic scale. The left side of the map shows how persons (students) distributed based on their ability (SRA) ("#" represents four persons and "." one person). The persons at the top are those getting high SRA, those at the bottom getting low SRA. On the right side of the map, the items scattered according to their difficulty levels. Items on the top are at a high level of difficulty and items at the bottom are at the low difficulty level. In addition, the CSR framework is added in the right section for easier reference. As illustrated, the Wright map of the SRA test demonstrates a good distribution of both students and test items. Students were approximately normally distributed based on their SRA. The majority of the test items were at the typical difficulty level (indicated by M on the right), which matched the typical ability level of the students (indicated by M on the left). The good match between the item concentration pattern and the student ability concentration pattern indicated "optimized measurement precision in this instrument's construction" (Juttner et al., 2013), confirming good construct validity of the developed SRA test (Aryadoust, 2009).

Also, as noted, the item difficulty level, in general, corresponded to the CSR framework. This further validated the SRA test constructed. There were some items (P23, P25 in particular) whose actual difficulty level deviated from the one prescribed (as shown in Table 4), calling for further modification and improvement of the SRA test.

https://doi.org/10.33225/jbse/20.19.261

Figure 2
Wright map (person-item map) of the SRA test (N=593, 25 items)



According to the Rasch modeling results elaborated above, the SRA test was proved to be reliable and valid, and the 6-level CSR framework was appropriate and applicable for distinguishing students' SRA.

Discussion

The present research has developed an integrated analytical framework and assessment tool of students' SRA that emphasizes the use of evidence in reasoning based on literature review, expert validation through a Delphi study, and empirical validation using multiple statistical methods (i.e., the correlation between the developed SRA test and LCTSR; and Rasch modeling). Data analysis results affirmed the reliability and validity of the SRA framework and test that particularly focus on capturing evidence from contextual information in solving science-related problems. Science teachers and other science educators can readily adapt and apply the SRA assessment to analyze and diagnose student performance in educational practices that focus on SRA in and for science learning.

https://doi.org/10.33225/jbse/20.19.261

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

On the other hand, as reflected by the data, some items of the SRA test (P23 and P25 in particular) need to be reconsidered. According to the Rasch analysis, the perceived difficulty level of P23 did not match with the level prescribed by the CSR framework. Even though P23 was designed to be a question of low CSR, only a few students answered it correctly. The interview data gathered in the pilot study suggested it was because of the inappropriate wording of the contextual information (that would be transformed into evidence) that prevented the students from providing the correct response. This observation pointed out the inadequacy of expert validation of test design. For improving the readability and comprehensibility of test items, the test design should also take opinions of the student as the real test-takers as reference. P24 and P25 also have the problem of ambiguous expression based on the post-test interview and discussion data. To further improve the soundness of the SRA test, these items need to be reinvented or could be eliminated.

Evaluating and diagnosing student performance with a legitimate assessment is the very first step to improve their SRA for science learning. With the assessment results, teachers shall reflect upon their teaching practices in an analytical, sustained, and a critical way for enhancing student performance (McNeill & Krajcik, 2011). Besides quantitative, objective evaluation, teachers may also further recognize and discern students' SRA via qualitative methods such as analyzing student interviews, writing assignments (Keys, 1994), or showcasing and eliciting scientific reasoning in the science classroom (Furtak et al., 2010). Engaging students in reasoning processes as the scientists do by providing evidence-based guidance would foster their understanding of science, enhance their skills of reasoning, and help them become rational when encountering science-related problems in real life (Driver et al., 1994).

In addition to applying the SRA test as summative assessment to help identify areas for improvement and further action in teaching and learning, teachers are also advised to adopt the CSR framework in-classroom observation of students' discourses as they interact with the teacher, peers, and the scientific phenomenon or problem as formative assessment to enable assessment for learning. Such real-time contextualized evaluation and feedback enabled by the CSR framework can help empower "teaching by inquiry" (Gerber et al., 2001) and the scientific practices (e.g. reasoning based on evidence, communicating and arguing with peers with evidence, solving contextualized problems, or/and trying and doing experiments or surveys, etc.) highly encouraged by NGSS (2013) to enhance students' scientific reasoning abilities (Gerber et al., 2001; Johnson & Lawson, 1998).

Conclusions and Implications

In this research, the SRA assessment developed based on both qualitative and quantitative methods, though proved valid and reliable to apply further, does have some limitations. Firstly, adopting the paper-pencil test, the SRA test did not create authentic contexts where scientists solve real-world problems. The processes of doing science, including raising questions, forming hypotheses, conducting experiments, obtaining data, and providing explanations, in which scientific reasoning plays a significant role, were hardly elicited. Furthermore, as the nature of scientific reasoning is closely related to the nature of science (NOS) (which is a key element in science curricula), the assessment and development of SRA in students should be deeply embedded in the processes of the functioning of science, the generation and testing of scientific knowledge, and the working of real scientists (McComas, 2011; Taber, 2006). In the present study, though the test items were designed to simulate authentic practices of science as much as possible, in-depth understanding of science and scientific reasoning still lacked. To collect and interpret holistic, process-oriented data on SRA, qualitative approaches are in need to help triangulate data. Considering the richness of scientific reasoning in communicative interactions in the science classroom, qualitative methods such as discourse analysis and process evaluation could be used combining with the quantitative methods in future investigations.

Moreover, the selection of subjects for the quantitative method relied on convenient sampling due to limited resources, which might negatively impact the generalizability of the findings. In the future, random selection is to be performed to produce more convincing results. And the data of student samples at different stages of schooling should be put into Rasch modeling to strengthen the test design further. Also, as the basis of the SRA test, the CSR framework has been inspected and validated by experts. As only a small group of experts were involved, their opinions might be biased or limited. Further iteration will be implemented to improve the assessment framework and the tool.

Acknowledgements

The research is the result of a project funded by "Zhejiang Office for Philosophy and Social Sciences Development and Planning" (Project No.: 20NDQN274YB), and Minister of Education (MOE) Key Research Institute of Humanities and Social Sciences, China (PRC). Also, it is sponsored by K. C. Wong Magna Fund in Ningbo University.

https://doi.org/10.33225/jbse/20.19.261

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

References

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

Aryadoust, S. V. (2009). Mapping Rasch-based measurement onto the argument-based validity framework. *Rasch Measurement Transactions, 23*(1), 1192-1193. https://www.rasch.org/rmt/rmt231f.htm

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

Belland, B. R., Glazewski, K. D., & Richardson, J. C. (2008). A scaffolding framework to support the construction of evidence-based arguments among middle school students. *Educational Technology Research and Development, 56*(4), 401-422. https://doi.org/10.1007/s11423-007-9074-1

Bolger, F., & Wright, G. (2011). Improving the Delphi process: lessons from social psychological research. *Technological Forecasting and Social Change, 78*(9), 1500-1513. https://doi.org/10.1016/j.techfore.2011.07.007

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

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE Life Sciences Education, 15*(4), rm4. https://dx.doi.org/10.1187%2Fcbe.16-04-0148

Brown, N. J. S., Furtak, E. M., Timms, M. J., Nagashima, S. O., & Wilson, M. (2010a). The evidence-based reasoning framework: assessing scientific reasoning. *Educational Assessment, 15*(3-4), 123-141. https://doi.org/10.1080/10627197.2010.530551

Brown, N. J. S., Nagashima, S. O., Fu, A., Timms, M. J., & Wilson, M. (2010b). A framework for analyzing scientific reasoning in assessments. *Educational Assessment, 15*(3-4), 142-175. https://doi.org/10.1080/10627197.2010.530562

Bybee, R. W., Fensham, P. J., & Laurie, R. (2009). Scientific literacy and contexts in PISA 2006 science. *Journal of Research in Science Teaching, 46*(8), 862-864. https://doi.org/10.1002/tea.20332

Choi, J. I., & Hannafin, M. (1997). The effects of instructional context and reasoning complexity on mathematics problem-solving. *Educational Technology Research and Development, 45*(3), 43-55. https://doi.org/10.1007/BF02299728

Churchill, R. P. (1990). *Instructor's manual to accompany logic, an introduction.* St. Martin's Press.

Coletta, V. P., Phillips, J. A., & Steinert, J. J. (2007). Why you should measure your students' reasoning ability. *The Physics Teacher, 45*(4), 235-238. https://doi.org/10.1119/1.2715422

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

Dohn, N. B. (2007). Knowledge and skills for PISA – Assessing the assessment. *Journal of Philosophy of Education, 41*(1), 1-16. https://doi.org/10.1111/j.1467-9752.2007.00542.x

Dolan, E. L., & Grady, J. (2010). Recognizing students' scientific reasoning: A tool for categorizing complexity of reasoning during teaching by inquiry. *Journal of Science Teacher Education, 21*(1), 31-55. https://doi.org/10.1007/s10972-009-9154-7

Driver, R., Asoko, H., Leach, J., Mortimer, E., & Scott, P. (1994). Constructing scientific knowledge in the classroom. *Educational Researcher, 23*(7), 5-12. https://doi.org/10.3102/0013189X023007005

Duschl, R. A. (2003). Assessment of inquiry. In J. M. Atkin & J. Coffey (Eds.), *Everyday assessment in the science classroom* (pp. 41-59). NSTA Press.

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

Fensham, P. J. (2009). Real world contexts in PISA science: Implications for context - based science education. *Journal of research in science teaching, 46*(8), 884-896. https://doi.org/10.1002/tea.20334

Furtak, E. M., Hardy, I., Beinbrech, C., Shavelson, R. J., & Shemwell, J. T. (2010). A framework for analyzing evidence-based reasoning in science classroom discourse. *Educational Assessment, 15*(3-4), 175-196. https://doi.org/10.1080/10627197.2010.530553

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

Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review, 16*(3), 235-266. https://doi.org/10.1023/B:EDPR.0000034022.16470.f3

Isaksen, S. G. (1998). *A review of brainstorming research: Six critical issues for inquiry*. Creative Research Unit, Creative Problem Solving Group-Buffalo. http://petkoivanov.com/wp-content/uploads/2015/10/302-Brainstorm.pdf

Jimenez-Aleixandre, M. P., Rodriguez, A. B., & Duschl, R. A. (2000). "Doing the lesson""Doing science": Argument in high school genetics. *Science Education, 84,* 757-792. https://doi.org/10.1002/1098-237X(200011)84:6<757::AID-SCE5>3.0.CO;2-F

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. http://doi.org/10.1080/09500693.2019.1594442 Johnson, M. A., & Lawson, A. E. (1998). What are the relative effects of reasoning ability and prior knowledge on biology achievement in expository and inquiry classes?. *Journal of Research in Science Teaching, 35*(1), 89-103. https://doi.org/10.1002/(SICI)1098-2736(199801)35:1<89::AID-TEA6>3.0.CO;2-J

Juttner, M., Boone, W., Park, S., & Neuhaus, B. J. (2013). Development and use of a test instrument to measure biology teachers' content knowledge (CK) and pedagogical content knowledge (PCK). *Educational Assessment, Evaluation and Accountability, 25*(1), 45-67. https://doi.org/10.1007/s11092-013-9157-y

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

https://doi.org/10.33225/jbse/20.19.261

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Kanari, Z., & Millar, R. (2004). Reasoning from data: How students collect and interpret data in science investigations. *Journal of Research in Science Teaching, 41*(7), 748-769. https://doi.org/10.1002/tea.20020

Keys, C. W. (1994). The development of scientific reasoning skills in conjunction with collaborative writing assignments: An interpretive study of six ninth-grade students. *Journal of Research in Science Teaching, 31*(9), 1003-1022. https://doi.org/10.1002/tea.3660310912

Kind, P. M., & 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

Kind, P. M. (2013). Establishing Assessment Scales Using a Novel Disciplinary Rationale for Scientific Reasoning. *Journal of Research in Science Teaching, 50*(5), 530-560 https://doi.org/10.1002/tea.21086

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*(1), 1-48. https://doi.org/10.1016/0364-0213(88)90007-9

Kuhn, D., Garcia-Mila, M., Zohar, A., Andersen, C., White, S., Klahr, D., & Carver, S. (1995). Strategies of Knowledge Acquisition. *Monographs of the Society for Research in Child Development, 60*(4), 1-157. https://doi.org/10.2307/1166059

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?!. *Intelligence, 14*(4), 389-433. https://doi.org/10.1016/S0160-2896(05)80012-1

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. https://doi.org/10.1007/s10763-004-3224-2

Lee, C. Q., & She, H. C. (2010). Facilitating students' conceptual change and scientific reasoning involving the unit of combustion. *Research in Science Education, 40*(4), 479-504. https://doi.org/10.1007/s11165-009-9130-4

Leron, U. (1985). A direct approach to indirect proofs. *Educational Studies in Mathematics, 16*(3), 321-325. https://doi.org/10.1007/BF00776741

Linacre, J. M. (2006). *A user's guide to WINSTEPS/MINISTEP: Rasch-model computer programs.* Winsteps.com.

Linacre, J. M. (2011). *A user's guide to WINSTEPS/MINISTEP: Rasch-model computer programs.* Winsteps.com.

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

McComas, W. F. (2011). *The history of science and the future of science education: Adapting Historical Knowledge Production to the Classroom.* Sense Publishers.

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*. Pearson.

Minner, D. D., Levy, A. J., & Century, J. (2010). Inquiry-based science instruction – what is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching, 47*(4), 474-496.

Moshman, D. (1998). Cognitive development beyond childhood. In D. Kuhn & R. Siegler (Eds.), W. Damon (Series Ed.), *Handbook of child psychology* (5th Ed.). Vol. 2. *Cognition, perception, and language* (pp. 947-978). Wiley. https://digitalcommons.unl.edu/edpsychpapers/48

National Research Council. (2012). *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. The National Academies Press.

NGSS Lead States. (2013). *Next generation science standards: For states, by states.* The National Academies Press.

OECD. (2006). PISA released items - science. http://www.oecd.org/pisa/38709385.pdf

OECD. (2015). Science framework. https://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Science%20Framework%20.pdf

Osborn, A. F. (1963). Applied imagination: Principles and procedures of creative problem-solving (3rd rev. ed.). Scribner.

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*(10), 994-1020. https://doi.org/10.1002/tea.20035

Randall, J. (2010). Using confirmatory factor analysis and the Rasch model to assess measurement invariance in a high stakes reading assessment. *Applied Measurement in Education, 23*(3), 286-306. https://doi.org/10.1080/08957347.2010.486289

Rowe, G., & Wright, G. (2001). Expert opinions in forecasting: The role of the Delphi technique. In *Principles of forecasting* (pp. 125-144). Springer US.

Sadler, T. D., & Zeidler, D. L. (2009). Scientific literacy, PISA, and socioscientific discourse: Assessment for progressive aims of science education. *Journal of Research in Science Teaching, 46*(8), 909-921. https://doi.org/10.1002/tea.20327

Salgado, F. A. (2016). Investigating the impact of context on students' performance. In White, B.. Chinnappan, M. & Trenholm, S. (Eds.), *Opening up mathematics education research (Proceedings of the 39th annual conference of the Mathematics Education Research Group of Australasia)* (pp.102-109). Mathematics Education Research Group of Australasia (MERGA). https://files.eric.ed.gov/fulltext/ED572407.pdf

Sandoval, W. A., & Millwood, K. A. (2005). The quality of students' use of evidence in written scientific explanations. *Cognition and Instruction, 23*(1), 23-55. https://doi.org/10.1207/s1532690xci2301_2

Sondergeld, T. A., & Johnson, C. C. (2014). Using Rasch measurement for the development and use of affective assessments in science education research. *Science Education, 98*(4), 581-613. https://doi.org/10.1002/sce.21118

Taber, K. S. (2006). Beyond constructivism: The progressive research programme into learning science. *Studies in Science Education, 42*(1), 125-184. https://doi.org/10.1080/03057260608560222

Tamassia, C., & Schleicher, A. (2002). Sample Tasks from the PISA 2000 Assessment: Reading, Mathematical and Scientific Literacy. http://www.oecd.org/education/school/programmeforinternationalstudentassessmentpisa/33692744.pdf

https://doi.org/10.33225/jbse/20.19.261

Thompson, E. D., Bowling, B. V., & Markle, R. E. (2017). Predicting student success in a major's introductory biology course via logistic regression analysis of scientific reasoning ability and mathematics scores. *Research in Science Education, 48*(1), 151-163. https://doi.org/10.1007/s11165-016-9563-5

Toulmin, S. E. (2003). *The uses of argument.* Cambridge University Press. *(Original work published 1958)*

Van Eemeren, F. H., Grootendorst, R., & Snoeck Henkemans, A. F. (2002). Argumentation: Analysis, evaluation, presentation. Lawrence Erlbaum Associates.

Wei, S., Liu, X., & Jia, Y. (2014). Using Rasch measurement to validate the instrument of students' understanding of models in science (sums). *International Journal of Science & Mathematics Education, 12*(5), 1067-1082. https://doi.org/10.1007/s10763-013-9459-z

Yanto, B. E., Subali, B., & Suyanto, S. (2019). Measurement instrument of scientific reasoning test for biology education students. *International Journal of Instruction, 12*(1), 1383-1398. https://doi.org/10.29333/iji.2019.12188a

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

Received: *January 02, 2020* Accepted: *April 04, 2020*

Cite as: Luo, M., Wang, Z., Sun, D., Wan, Z., & Zhu, L. (2020). Evaluating scientific reasoning ability: The design and validation of an assessment with a focus on reasoning and the use of evidence. *Journal of Baltic Science Education*, *19*(2), 261-275. https://doi.org/10.33225/jbse/20.19.261

Ma Luo PhD, Lecturer, College of Teacher Education, Ningbo University, Ningbo, China.
E-mail: rome9009@hotmail.com
ORCID: https://orcid.org/0000-0003-2105-967X

Zuhao Wang PhD, Professor, College of Teacher Education, East China Normal University, Shanghai, China
E-mail: wangzuhao@126.com

Daner Sun *(Corresponding author)* PhD, Assistant Professor, Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong SAR, China.
E-mail: dsun@eduhk.hk
Website: https://repository.eduhk.hk/en/persons/daner-sun
ORCID: https://orcid.org/0000-0002-4516-0505

Zhihong Wan PhD, Assistant Professor, Department of Curriculum and Instruction, The Education University of Hong Kong, Hong Kong SAR, China.
E-mail: wanzh@eduhk.hk
Website: https://repository.eduhk.hk/en/persons/zhihong-wan
ORCID: https://orcid.org/0000-0002-8163-5862

Liying Zhu PhD, Lecturer, College of Teacher Education, Zhejiang Normal University, Jinhua, China.
E-mail: zhuly2000@163.com
ORCID: https://orcid.org/0000-0001-9251-0304