ISEJ Indonesian Science Education Journal

ISSN Online: 2716-3350
https://siducat.org/index.php/isej/
Published by Yayasan Darussalam Bengkulu

OPEN ACCESS ARTICLE

Development of a Higher-Order Thinking Skills Assessment Instrument Using Socioscientific Issue Contexts on Chemical Solutions

Ervita Eka Rosawati
Science Education, Universitas Islam Negeri Sunan Ampel Surabaya, Surabaya, Indonesia
E-mail Corresponding: ervita.rosawati@uinsa.ac.id

Received: November 24, 2025	Revised: December 27, 2025	Accepted: January 21, 2026

Abstract
This study aimed to develop and validate a higher-order thinking skills (HOTS) assessment instrument using socioscientific issue (SSI) contexts on chemical solution topics for senior high school students. The study employed an adapted research and development design consisting of five stages: preliminary analysis, item development, expert validation, limited trials, and field implementation. The instrument consisted of 40 items, comprising 35 multiple-choice items and five essay items, embedded in nine SSI contexts related to buffer solutions, acid-base titration, and colloids. The participants were 432 twelfth-grade science students from ten public senior high schools across districts and cities in East Java, Indonesia. Data were analyzed using expert judgment, Pearson product-moment correlation, Cronbach's Alpha, Cohen's Kappa, and the Partial Credit Model within the Rasch measurement framework. The findings showed that the instrument had strong content validity, with an overall expert validation score of 83.91%. The instrument also demonstrated good internal consistency, as indicated by a Cronbach's Alpha of 0.81, item reliability of 0.98, person reliability of 0.82, item separation of 7.88, and person separation of 2.10. The essay scoring showed strong inter-rater agreement, with a Kappa coefficient of 0.841. Rasch analysis indicated that most items fit the model, although several items required further review due to relatively high outfit and infit values or low point-measure correlation. The findings suggest that the developed SSI-based HOTS assessment instrument is psychometrically feasible for measuring students' higher-order thinking in chemistry learning. This study implies that contextualized assessment through socioscientific issues can support more authentic measurement of students' analytical, evaluative, and creative thinking in chemical solution topics.

Keywords: Assessment Instrument; Chemical Solutions; Higher-Order Thinking Skills; Rasch Model; Socioscientific Issues

INTRODUCTION

The use of socioscientific issues (SSI) in science education has gained attention in recent decades as an important area of study. SSI helps students contextualize their scientific knowledge, develop critical thinking, multi-perspectival reasoning, argumentation skills, and encourage moral and ethical thinking regarding science in society (Chowning et al., 2012; Dawson & Venville, 2022; Sadler et al., 2007). These studies highlight the importance of SSI in achieving scientific literacy (Lee et al., 2013; Ratcliffe & Grace, 2003; Sadler, 2011). A socioscientific issue is a controversial, ill-structured, and dilemmatic issue related to science (Sadler, 2004), requiring moral reasoning or evaluation of ethical-related issues to make decisions. Although SSIs do not have clear solutions, they still have some reasonable solutions based on scientific principles, data, and theory, while also requiring consideration of social factors such as politics, ethics, and economics (Sadler, 2011). SSI is purposively used in science learning to encourage conversation, debate, and discussion. It demands reasoning based on scientific evidence, is interesting and

meaningful for students, and offers a context for comprehending scientific material (Sadler, 2004; Zeidler et al., 2019). Studies have shown that SSI can enhance a broad range of abilities, lerincluding argumentation skills, critical thinking, and comprehension of scientific concepts.

However, there is no research on higher order thinking skills (HOTS) assessment instruments that use SSI as a context, as most studies have focused on SSI in the teaching and learning process (Qamariyah et al., 2021; Rahayu & Alsulami, 2024; Zohar & Nemet, 2002). Research related to the use of SSI in the development of assessment instruments is still limited. Most SSI applications remain focused on learning processes rather than assessment development. Previous studies also show that the use of HOTS questions in high school chemistry is still relatively low, with assessment instruments dominated by lower-order thinking skills (Iskandar & Senam, 2015; Nadhifah, 2019). However, the use of SSI as the context in HOTS assessment instruments is still rare.

Developing higher-order thinking abilities is a crucial part of teaching and learning. According to Anderson and Krathwohl (2001), higher-order thinking skills include analyze, evaluate, and create. Therefore, HOTS is a thinking skill that not only relies on the ability to remember, but also higher-level abilities such as analyzing, evaluating, and creating. HOTS also extends beyond the Revised Bloom taxonomy, as it requires deep conceptual understanding, unfamiliar contexts, metacognitive dimensions, and critical thinking (Anderson & Krathwohl, 2001; Zohar & Nemet, 2002). Assessment in education has a large effect on learning, including how teachers determine the content to be taught and how the learning process will take place (Baird et al., 2017). Assessment instruments used by teachers can determine the success of developing higher-order thinking skills in chemistry learning (Fensham & Bellocchi, 2013). If the assessment instrument requires students to think at a higher level, then students will have a greater opportunity to develop these abilities (FitzPatrick & Schulz, 2015). However, teachers still experience difficulties in developing assessment instruments that measure students' higher-order thinking skills (Dahlan et al., 2020; Driana & Ernawati, 2019; FitzPatrick & Schulz, 2015).

The chemical solution materials used in this study were buffer solutions, acid-base titrations, and colloids. These topics were selected because they are relevant to chemistry learning and closely connected to real-life socioscientific contexts. To obtain accurate test analysis results, an appropriate assessment analysis method is needed. One method that can be used is the Rasch model, which evaluates students' responses based on item difficulty and student ability (Engelhard Jr., 2013). Rasch analysis is widely regarded as a strong measurement method in educational assessment (Bond & Fox, 2012; Boone et al., 2014).

This study aimed to develop a higher order thinking skills assessment instrument using socioscientific issue contexts in chemical solution topics. The developed instrument was intended to measure students' HOTS and examine its feasibility using Rasch model analysis.

RESEARCH QUESTION

This study aimed to develop a Higher-Order Thinking Skills (HOTS) assessment instrument using Socioscientific Issue contexts on chemical solutions and examine its validity and reliability. The research questions are: (1) What are the validity and reliability of a Higher-Order Thinking Skills assessment instrument on chemical solutions developed using an SSI context? (2) Is the developed HOTS assessment instrument feasible for measuring students' higher-order thinking skills on chemical solutions?

METHODS

This study involved the development of a Higher-Order Thinking Skills (HOTS) assessment instrument using a Socioscientific Issue (SSI) context on chemical solutions. The development consisted of five steps: preliminary study, item development, expert judgment, first trial, and second trial (Borg & Gall, 2003). The developed instrument was analyzed using the Partial Credit Rasch Model to assess its feasibility.

In the development stages, the validity of the items was determined using the Pearson

Bivariate correlation formula (Pearson Product Moment), and the reliability of the test instrument was tested using Cronbach's Alpha Coefficient with SPSS 25.0. Essay questions were analyzed by 2 raters, and inter-rater reliability was measured using Cohen's Kappa Coefficient. After implementation, the Rasch Model was used to analyze the instrument.

Participants

The participants of this study were twelfth-grade science students from 10 state high schools across 5 districts and 4 cities in East Java, Indonesia, with a total of 432 students (175 male, 257 female). The chemistry topics assessed in this study were buffer solutions, acid-base titrations, and colloids, which had been studied by the students in the second semester of grade XI. East Java was selected as the study setting because it was considered representative for the implementation of the developed HOTS assessment instrument.

Table 1. Distribution of the participants

Distric/City	Participants
Sidoarjo	62
Gresik	33
Mojokerto	44
Pasuruan	38
Jombang	38
City of Surabaya	77
City of Malang	60
City of Mojokerto	43
City of Pasuruan	37
Total	432

Item Development

The assessment instrument was developed to measure students' higher-order thinking skills in chemical solution topics, namely buffer solutions, acid-base titrations, and colloids, using socioscientific issue contexts. The instrument consisted of 40 items (35 multiple-choice and 5 essay questions) with 9 socioscientific issue texts. Based on the cognitive process dimensions, 18 items measured Analyzing (C4), 14 items measured Evaluating (C5), and 8 items measured Creating (C6).

The items were validated by four expert validators, consisting of two chemistry lecturers and two chemistry teachers. A legibility test was also conducted to assess language, concept suitability, clarity of graphs/tables/images, and layout appearance. The higher-order thinking operational verbs used in this assessment instrument are classified in Table 2.

Table 2. HOTS operational verbs used in the assessment instrument

Knowledge dimension	Cognitive Process Dimension		
	C4 analyzing	C5 evaluating	C6 creating
Factual knowledge	-	-	-
Conceptual knowledge	analyzing, differentiating	checking criticizing	-
Procedural knowledge	analyzing	checking criticizing	planning
Metacognitive knowledge	-	-	creating

Multiple-choice items were scored 1 for correct answers and 0 for incorrect answers, while essay responses were scored using specified ranges. Essay answers were assessed by two raters, a researcher and a chemistry teacher. The mapping of questions and the context of

socioscientific issues in each material is shown in Table 3.

Table 3. The number of questions and the context of socioscientific issues in each material

Material	The Cognitive Process Dimension	Socioscientific issues context	Number of questions	
			Multiple choice	Essay
Buffer solution	C4, C5, C6	Leather industrial waste in Malang	6	2
	C4, C6	[Facts or hoaxes] about the benefits of consuming alkaline foods to ward off the Covid-19 virus	4	-
	C4, C5, C6	The dilemma of cancer treatment	3	1
Acid-base titration	C4, C5	B3 waste emergency at hospitals in East Java	5	-
	C4, C5, C6	Controversy over the use of ibuprofen for the treatment of Covid-19	4	-
	C4	Illegal dolomite mines	3	-
Colloid	C5, C6		2	1
	C4, C5	The use of soap vs hand sanitizer to fight the Covid-19 virus	2	-
	C4, C6	Indonesia's “killer” fog	1	1
	C4, C5	The dilemma of using formula milk	5	-
Total			**35**	**5**

Data Collection

Participants were selected using convenience sampling and had previously learned the chemical solution topics assessed in this study, namely buffer solutions, acid-base titrations, and colloids. Participation was voluntary, as stated in the participant consent form. The assessment instrument was administered via Google Form under the supervision of teachers and researchers.

The instrument was tested in two trials. The first trial involved 70 students, and the second involved 181 students. The Cronbach Alpha values were 0.86 and 0.82, respectively. Based on the validity analysis, invalid items were revised after each trial, resulting in a final instrument consisting of 40 items. The final instrument was then administered to 432 students.

Data Analysis

The developed items were validated by four experts based on substance, construction, and language aspects. The instrument was then analyzed using the Partial Credit Model in the Rasch approach with WINSTEPS and Excel. Rasch analysis was used to examine item validity, reliability, item and person separation, item and person reliability, Cronbach’s alpha, unidimensionality, outfit MNSQ, infit value, and Pt-measure correlation to evaluate the feasibility of the developed HOTS assessment instrument.

RESULT AND DISCUSSION

Reliability

The reliability of the assessment instrument is divided into item reliability, Cronbach's Alpha value, and separation. The HOTS assessment tool's reliability is displayed in Table 4.

Table 4. The reliability of the HOTS assessment instrument

Parameter	Item	Person
Reliability	0.98	0.82
Separation	7.88	2.10
Cronbach's Alpha	0.81	

The Cronbach Alpha value of 0.81 suggests that the items were developed with a high degree of reliability. According to Bond and Fox (2012), an assessment obtaining a Cronbach alpha value of 0.7 suggests a level of reliability that is acceptable, whereas an alpha value of 0.8 is good and 0.9 is very good. The reliability of the item is 0.98 indicating that this item is highly trusted. Additionally, the value of item separation is greater than three, making it very good (Bond & Fox, 2012; Duncan et al., 2003).

The test questions in the form of essays were analyzed by 2 raters, so the developed test instrument was also tested for inter-rater reliability to measure internal consistency. The results showed a Kappa value of 0.841, indicating strong inter-rater agreement. Table 5. shows the results of calculating inter-rater reliability at the implementation stage.

Table 5. Inter-rater reliability test results

	Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement Kappa	0.841	0.040	31.720	0.000
N of Valid Cases	87			

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

Validity

Instruments that have been developed were consulted with experts to determine the feasibility of the resulting format and the suitability of the learning material. Overall, the HOTS assessment instrument in the SSI context has a validity value of 83.91%, indicating that the assessment instrument was very valid to implement.

Several aspects were measured in this validation process. The substance aspect had 84.22%, representing very good validity, including the suitability of questions with indicators, competencies, answer keys, and the homogeneity and logic of the answer choices. The construction aspect had 83.04%, also representing very good validity, including clarity of item formulation, absence of answer key clues, and avoidance of negative statements. The language aspect also indicated very good validity, covering the use of appropriate and communicative language.

After implementation, the validity of the instrument was further analyzed through Rasch modeling by examining item characteristics, as presented in Table 6. Item fit statistics were used to determine whether the items fit the Rasch model. According to Bond and Fox (2012), MNSQ infit and outfit values within the range of 0.5 < MNSQ < 1.5 meet the acceptable criteria.

Table 6. Item fit statistic of HOTS assessment instrument

Item	Code	Item logit	Standard Error	Infit MNSQ	Outfit MNSQ	Pt-measure correlation
Q22	A	1.77	0.16	0.94	0.96	0.30
Q35	A	1.15	0.13	0.96	1.03	0.28
Q6	A	1.10	0.13	1.02	1.07	0.19
Q23	A	1.05	0.13	0.97	1.02	0.27
Q29	A	1.05	0.13	1.05	1.11	0.13
Q37	A	1.02	0.12	0.98	1.03	0.26
Q12	A	1.01	0.12	0.99	1.02	0.26
Q32	A	1.01	0.12	0.95	0.97	0.33
Q19	A	0.95	0.12	0.99	1.04	0.24
Q17	A	0.93	0.12	1.02	1.09	0.19
Q1	A	0.90	0.12	1.04	1.08	0.17

Item	Code	Item logit	Standard Error	Infit MNSQ	Outfit MNSQ	Pt-measure correlation
Q14	A	0.87	0.12	0.98	1.00	0.27
Q5	A	0.86	0.12	1.03	1.06	0.19
Q11	A	0.80	0.12	1.07	1.13	0.10
Q24	B	0.75	0.12	0.96	0.96	0.32
Q34	B	-0.16	0.04	1.08	1.12	0.37
Q8	B	-0.20	0.04	1.24	1.25	0.40
Q31	C	-0.24	0.04	1.14	1.15	0.35
Q16	C	-0.27	0.03	1.52	1.64	0.40
Q7	A	-0.32	0.03	1.58	1.80	0.43
Q33	A	-0.37	0.10	0.93	0.93	0.37
Q26	A	-0.40	0.10	0.87	0.85	0.47
Q9	A	-0.41	0.10	0.92	0.93	0.38
Q27	A	-0.43	0.10	0.89	0.87	0.44
Q13	A	-0.51	0.10	0.86	0.84	0.48
Q28	A	-0.54	0.10	0.89	0.87	0.44
Q30	A	-0.54	0.10	0.90	0.88	0.42
Q21	A	-0.56	0.10	0.87	0.85	0.47
Q18	A	-0.59	0.10	0.94	0.95	0.33
Q36	A	-0.59	0.10	0.92	0.90	0.38
Q25	A	-0.61	0.10	0.90	0.87	0.43
Q15	A	-0.66	0.10	0.93	0.93	0.36
Q39	A	-0.69	0.10	0.87	0.84	0.47
Q20	A	-0.78	0.10	0.91	0.88	0.40
Q40	A	-0.82	0.10	0.88	0.84	0.45
Q10	A	-0.94	0.10	0.94	0.92	0.33
Q3	A	-1.03	0.10	0.95	0.94	0.32
Q38	A	-1.13	0.10	0.88	0.82	0.45
Q2	A	-1.24	0.11	0.94	0.90	0.33
Q4	A	-1.24	0.11	0.92	0.87	0.36

A: Multiple choice items
B: Essay items (1-5 point)
C: Essay items (1-10 point)

Table 6 shows that the developed items generally met these requirements, indicating that the instrument was capable of measuring the intended constructs. In addition, the Pt-measure correlation values indicate acceptable item discrimination, and all items showed standard error values below 0.5, supporting the precision of the developed instrument

Construct validity was also examined through unidimensionality analysis. Based on Rasch analysis, the raw variance explained in this measurement was 32%, while the unexplained variance for all contrasts remained below 15%. Although one contrast showed an eigenvalue above 2, the overall results indicate that the instrument remained acceptable and was not substantially influenced by other dimensions.

Item difficulty analysis further showed that all items fell within the moderate difficulty category, ranging from -1.24 to 1.77 logits, as shown in Table 6. This indicates that the developed instrument contained items with an appropriate level of difficulty for the intended participants. The distribution of item difficulty is also illustrated in Figure 1.

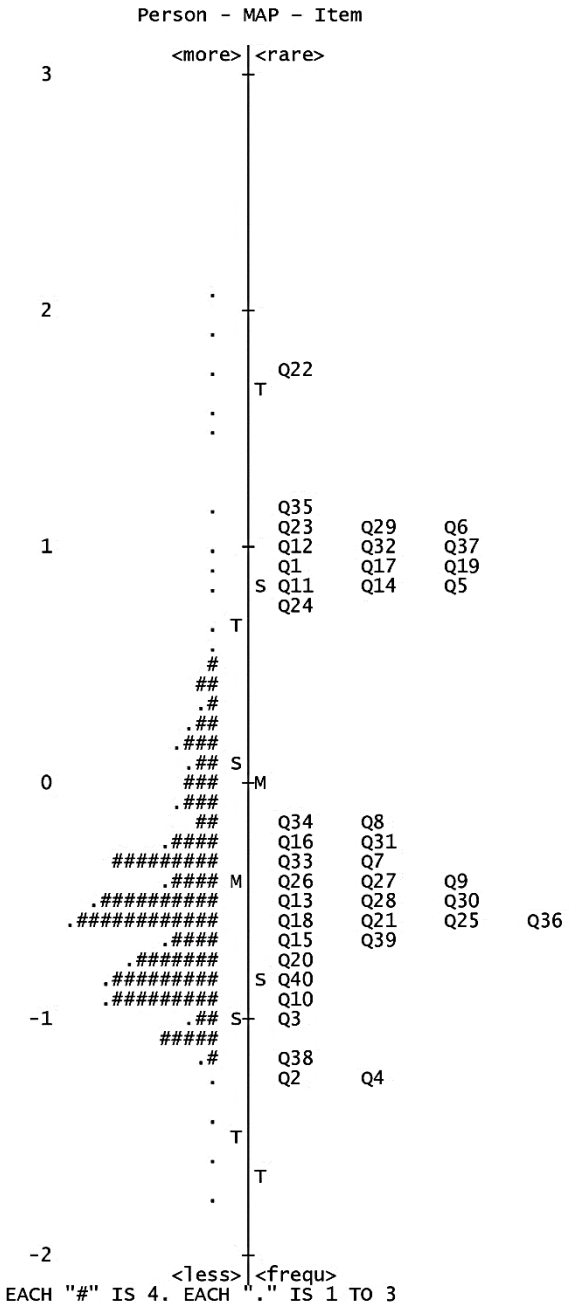

Figure 1. Wright Map of Items and Persons

Figure 1 illustrates the range of questions from easy to difficult. The list of questions is arranged with easy questions at the bottom and difficult questions at the top. These results suggest that the developed instrument contains a range of item difficulty levels appropriate for measuring higher-order thinking skills.

Discussion

The findings indicate that the developed HOTS assessment instrument using socioscientific issue (SSI) contexts on chemical solution topics demonstrates strong psychometric quality. The reliability analysis showed a Cronbach's Alpha value of 0.81, indicating good internal consistency according to Bond and Fox (2012). The item reliability value of 0.98 and item separation index of 7.88 further suggest that the developed instrument has excellent consistency and is capable of distinguishing item difficulty levels effectively. These findings indicate that the instrument was constructed with a stable measurement structure, making it appropriate for assessing higher-order thinking skills.

The strong inter-rater reliability result (Kappa = 0.841) also supports the consistency of scoring, particularly for essay items that involve subjective judgment. This suggests that the scoring rubric and scoring procedures were sufficiently clear to minimize variation between raters. Since HOTS assessment often includes open-ended responses requiring evaluative judgment, establishing strong inter-rater agreement is essential to ensure scoring credibility.

The expert validation results also confirmed the quality of the developed instrument. The overall validity score of 83.91% indicates that the assessment instrument is highly feasible for implementation. The high scores across substance, construction, and language aspects suggest that the instrument aligns well with the intended competencies, uses appropriate item construction principles, and presents content in a clear and communicative manner. This finding is particularly important because the development of HOTS assessment instruments remains a challenge for many teachers, as reported by Dahlan et al. (2020), Driana and Ernawati (2019), and Schulz and FitzPatrick (2016).

Further evidence of validity was demonstrated through Rasch analysis. Most items met the acceptable fit criteria based on infit and outfit MNSQ values, indicating that the items functioned consistently in measuring the intended construct. The Pt-measure correlation values also suggest acceptable discrimination among items, while the low standard error values indicate precise item measurement. These findings support the use of Rasch modeling as a robust approach for evaluating educational assessment instruments, consistent with Engelhard (2013), Bond and Fox (2012), and Boone et al. (2014).

The unidimensionality analysis also provides important evidence regarding construct validity. Although one contrast showed an eigenvalue above the ideal threshold, the explained variance of 32% and unexplained variance below 15% across contrasts suggest that the instrument remains sufficiently unidimensional for practical measurement purposes. This indicates that the developed assessment instrument primarily measures a common construct related to higher-order thinking skills, despite the complexity of combining multiple HOTS dimensions and SSI-based contexts.

The distribution of item difficulty levels further strengthens the quality of the instrument. The Wright Map and item logit values show that the developed questions span a range of difficulty levels, from easier to more challenging items. This distribution is important in HOTS assessment because higher-order thinking should not be measured through uniformly difficult or uniformly easy items, but through a balanced range that reflects different levels of cognitive demand. The inclusion of analyzing (C4), evaluating (C5), and creating (C6) dimensions across multiple chemical solution contexts also supports the comprehensiveness of the instrument.

The integration of socioscientific issue contexts represents an important contribution of this study. Previous studies have predominantly used SSI as a learning context rather than as a framework for assessment development. By embedding SSI into HOTS assessment items, this study extends the application of SSI beyond instructional practice and demonstrates its potential in assessment design. Because SSI contexts require students to consider scientific evidence alongside social, ethical, and real-world considerations, they provide a meaningful context for eliciting higher-order thinking processes.

Overall, the developed instrument contributes to the growing need for valid and reliable HOTS assessment tools in chemistry education. However, future studies may further refine the instrument by expanding the range of SSI contexts, strengthening dimensional structure, and adapting the instrument to other science disciplines.

CONCLUSION

This research has led to the development of a valid and reliable instrument to measure students' higher-order thinking skills (HOTS) using the SSI context in the form of 35 multiple choice questions and 5 essay questions with 9 SSI contexts. Through Rasch Modeling, the item and person reliability, item and person separation, and Alpha Cronbach value, as well as the item validity in the form of the infit value, outfit MNSQ, Pt-measure correlation value, and

unidimensionality, confirmed the validity and reliability of the test instrument. Potentially, this instrument can be used to measure students' higher-order thinking skills. This assessment instrument can also be adapted to other disciplines. The selection of the SSI context in the instrument development process is essential, so it is hoped that in future research the selection of the context will be given more attention and adapted to the relevance of the material. The form of the assessment instrument is also expected to be further developed so that it can measure the dimensions of students' higher-order thinking skills more deeply.

REFERENCE

Anderson, L. W., & Krathwohl, D. R. (2001). *A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives: Complete Edition*. Longman.

Baird, J.-A., Andrich, D., Hopfenbeck, T. N., & Stobart, G. (2017). Assessment and learning: Fields apart? *Assessment in Education: Principles, Policy & Practice*, *24*(3), 317–350. https://doi.org/10.1080/0969594X.2017.1319337

Bond, T. G., & Fox, C. M. (2012). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer. https://doi.org/10.1007/978-94-007-6857-4

Borg, W. R., & Gall, M. D. (2003). *Educational research: An introduction* (7th ed.). Pearson Education, Inc.

Chowning, J. T., Griswold, J. C., Kovarik, D. N., & Collins, L. J. (2012). Fostering Critical Thinking, Reasoning, and Argumentation Skills through Bioethics Education. *PLoS ONE*, *7*(5), e36791. https://doi.org/10.1371/journal.pone.0036791

Dahlan, Permana, L., & Oktariani, M. (2020). Teacher Competence and Difficulties in Constructing HOTS Instruments in Economics Subject. *Cakrawala Pendidikan*, *39*(1), 111–119.

Dawson, V., & Venville, G. (2022). Testing a methodology for the development of socioscientific issues to enhance middle school students' argumentation and reasoning. *Research in Science & Technological Education*, *40*(4), 499–514. https://doi.org/10.1080/02635143.2020.1830267

Driana, E., & Ernawati, E. (2019). TEACHERS' UNDERSTANDING AND PRACTICES IN ASSESSING HIGHER ORDER THINKING SKILLS AT PRIMARY SCHOOLS. *ACITYA Journal of Teaching & Education*, *1*(2), 110–118. https://doi.org/10.30650/ajte.v1i2.233

Duncan, P. W., Bode, R. K., Min Lai, S., & Perera, S. (2003). Rasch analysis of a new stroke-specific outcome scale: the stroke impact scale11No commercial party having a direct financial interest in the results of the research supporting this article has or will confer a benefit upon the author(s) or upon any organization with which the author(s) is/are associated. *Archives of Physical Medicine and Rehabilitation*, *84*(7), 950–963. https://doi.org/10.1016/S0003-9993(03)00035-2

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

Fensham, P. J., & Bellocchi, A. (2013). Higher order thinking in chemistry curriculum and its assessment. *Thinking Skills and Creativity*, *10*, 250–264. https://doi.org/https://doi.org/10.1016/j.tsc.2013.06.003

FitzPatrick, B., & Schulz, H. (2015). Do Curriculum Outcomes and Assessment Activities in Science Encourage Higher Order Thinking? *Canadian Journal of Science, Mathematics and Technology Education*, *15*(2), 136–154. https://doi.org/10.1080/14926156.2015.1014074

Iskandar, D., & Senam. (2015). Studi Kemampuan Guru Kimia SMA Lulusan UNY dalam Mengembangkan Soal UAS Berbasis HOTS. *Jurnal Inovasi Pendidikan IPA*, *1*(1), 86–101.

Lee, H.-B., Yoo, J., Choi, K., Kim, S.-W., Krajcik, J., & Herman, B. C. (2013). Socioscientific issues as a vehicle for promoting character and values for global citizens. *International Journal of Science Education*, *35*(12), 2079–2113.

Nadhifah, Y. (2019). *Analisis keterlaksanaan aspek Higher Order Thinking Skills (HOTS) pada soal-soal ujian mata pelajaran kimia kurikulum 2013 SMA di Kota Malang*.

Qamariyah, S. N., Rahayu, S., Fajaroh, F., & Alsulami, N. M. (2021). The Effect of Implementation of Inquiry-based Learning with Socio-scientific Issues on Students' Higher-Order Thinking Skills. *Journal of Science Learning*, *4*(3), 210–218.

Rahayu, S., & Alsulami, N. M. (2024). Assessing higher order thinking skills of the 21st century learners using socio-scientific issues as a context. *AIP Conference Proceedings*, *3106*(1). https://doi.org/10.1063/5.0214793

Ratcliffe, M., & Grace, M. (2003). *Science education for citizenship: Teaching socioscientific issues*. Open University Press.

Sadler, T. D. (2004). Informal reasoning regarding socioscientific issues: A critical review of research. *Journal of Research in Science Teaching*, *41*(5), 513–536.

Sadler, T. D. (2011). Situating socioscientific issues in classrooms as a means of achieving goals of science education. In T. D. Sadler (Ed.), *Socioscientific issues in the classroom: Teaching, learning and research* (pp. 1–9). Springer.

Sadler, T. D., Barab, S. A., & Scott, B. (2007). What do students gain by engaging in socioscientific inquiry? *Research in Science Education*, *37*, 371–391. https://doi.org/10.1007/s11165-006-9030-9

Schulz, H., & FitzPatrick, B. (2016). Teachers' understandings of critical and higher order thinking and what this means for their teaching and assessments. *Alberta Journal of Educational Research*, *62*(1), 61–86.

Zeidler, D. L., Herman, B. C., & Sadler, T. D. (2019). New directions in socioscientific issues research. *Disciplinary and Interdisciplinary Science Education Research*, *1*(1), 11. https://doi.org/10.1186/s43031-019-0008-7

Zohar, A., & Nemet, F. (2002). Fostering student's knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching*, *39*(1), 35–62.