Can. J. Sci. Math. Techn. Educ. (2026) 26:48
https://doi.org/10.1007/s42330-026-00470-9

UNIVERSITY OF TORONTO
OISE | ONTARIO INSTITUTE
FOR STUDIES IN EDUCATION

Assessing Socioscientific Argumentation: Exploration of Potential Threats to Validity Using the Assessment Triangle

Nina Minkley[1] · **Carola Garrecht**[2] · **Moritz Krell**[2]

Received: 26 March 2025 / Accepted: 25 February 2026

Abstract This paper explores the challenges associated with assessing socioscientific argumentation (SSA) using the assessment triangle as a framework. Our considerations point to the difficulty of validly capturing SSA as it depends on how it is measured (observation) and how the construct is inferred from the observation (interpretation). Therefore, we systematically varied observation and interpretation by implementing four dilemmas and two coding schemes. In total, 64 preservice teachers for biology processed two to four dilemmas with different contexts, and their argumentation was analysed using two similar coding schemes for structural complexity to investigate whether there are systematic differences in the inferred level of SSA, even if the coding schemes are similar. Our findings show that, on average, the participants achieved an approximately medium level of SSA in both schemes. A significant difference in participants' SSA levels when coded with the same scheme was only found between four of the 12 possible pairwise dilemma combinations. With regard to the relevance of interpretation, the participants achieved significantly higher levels in one coding scheme than in the other, regardless of the dilemma. These results indicate that the specific dilemma can but does not necessarily influence the level of SSA. Conversely, the choice of coding scheme has a major influence on the interpretation, which stresses the importance of carefully choosing an appropriate coding scheme.

Résumé Dans cet article, nous abordons les défis liés à l'évaluation de l'argumentation socioscien-

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42330-026-00470-9.

✉ Nina Minkley
Nina.Minkley@rub.de

Carola Garrecht
garrecht@leibniz-ipn.de

Moritz Krell
krell@leibniz-ipn.de

[1] Biology Education, Faculty of Biology and Biotechnology, Ruhr-Universität Bochum, Bochum, Germany

[2] IPN – Leibniz Institute for Science and Mathematics Education, Kiel, Germany

Published online: 08 April 2026

tifique (ASS) en utilisant le triangle d'évaluation comme cadre. Nos réflexions soulignent la difficulté de cerner de manière valide l'ASS, car celle-ci dépend de la façon dont elle est mesurée (observation) et de la manière dont le concept est déduit de l'observation (interprétation). Nous avons donc systématiquement varié l'observation et l'interprétation en mettant en œuvre quatre dilemmes et deux schémas de codage. En tout, 64 enseignants en formation initiale en biologie se sont penchés sur deux à quatre dilemmes présentant des contextes différents, et nous avons analysé leur argumentation à l'aide de deux schémas de codage comparables pour la complexité structurelle afin de déterminer s'il existe des différences systématiques dans le degré d'ASS déduit, même lorsque les schémas de codage sont similaires. Nos recherches montrent qu'en moyenne, les participants ont atteint un niveau d'ASS relativement moyen dans les deux schémas. Une différence notable dans les résultats d'ASS des personnes testées avec le même schéma n'a été observée que dans quatre des douze combinaisons de dilemmes possibles. En ce qui concerne la pertinence de l'interprétation, les participants ont atteint des niveaux nettement plus élevés dans un schéma de codage plutôt que dans l'autre, quel que soit le dilemme. Ces résultats indiquent qu'un dilemme particulier peut influencer le degré d'ASS, mais pas forcément. À l'inverse, le choix du schéma de codage exerce une influence majeure sur l'interprétation, ce qui souligne l'importance de choisir avec soin un schéma de codage approprié.

Keywords Socioscientific argumentation · Assessment triangle · Validity

Introduction

An important goal of science education is to help students develop a scientific understanding that will empower them to participate in and contribute to science-related discussions of societal relevance (e.g. DeBoer, 2000). This includes engaging with socioscientific issues (i.e. 'social issues with conceptual or technological ties to science'; Sadler, 2004, p. 513), which — due to their controversial nature — require the evaluation and weighing of different and sometimes conflicting arguments and positions (Jiménez-Aleixandre & Erduran, 2008). This argumentative engagement with socioscientific issues, which is defined in this paper as socioscientific argumentation (SSA), can be described as complex and cognitively demanding, as multiple viewpoints need to be considered in the argumentation and there are usually no clear-cut solutions to the underlying problem (e.g.Bencze et al., 2020; Zeidler, 2014).

In consideration of the importance of SSA for science education, research on the (validity of) assessment of SSA is an essential objective of science education research (Nielsen, 2020). In the context of scientific reasoning competencies, Osborne et al. (2016) argues that the development of assessment instruments is pertinent to science education for several reasons: The utilisation of assessment instruments is conductive to the clear operationalisation and communication of constructs. Furthermore, these instruments have the potential to facilitate the integration of novel concepts into the implemented curriculum. This is due to the tendency of teachers to prioritise constructs that are the focus of (high stakes) tests. Additionally assessment items can be employed for teaching purposes in science classes. One of the challenges associated with SSA concerns the inconsistency of a person's SSA across different contexts (e.g. Topçu et al., 2010). As such, 'context dependencies' have been identified as a possible threat to valid assessments (Krell et al., 2014; Shavelson, 2013) and only few empirical studies on context dependencies of SSA are available to date; further research is needed on how SSA can be assessed validly and what potential threats to validity exist in the assessment of SSA.

Typically, the assessment of SSA in science education uses science-related scenarios (e.g. moral-ethical dilemmas) that require respondents to make or to evaluate a decision, which is then analyzed using a coding scheme to infer the respondents' level of SSA (e.g. Cetin et al., 2014; Topçu et al., 2010; Krell et al., 2024). This assessment approach harbors two major threats to validity: First, it can be

argued that the context of an assessment task (e.g. the specific moral-ethical dilemma) might influence respondents' SSA due to a possible lack of relevant content knowledge (e.g.Baytelman et al., 2020; Garrecht et al., 2021; Osborne et al., 2016; Sadler & Zeidler, 2005a; Topçu et al., 2010) or due to context-dependent affective features, such as motivation (see Zeidler, 2014). The context can, hence, influence respondents' performance (i.e. the specific type or quality of argumentation). From an assessment perspective, the influence of the context on performance poses a problem for the generalisation of test scores (Osborne et al., 2016) and thus for the development of reliable and valid assessment instruments (Shavelson, 2013). Second, various coding schemes for performance interpretation have been proposed (e.g. Reitschert et al., 2007; Sadler & Fowler, 2006). However, it remains unclear how these different coding schemes compare in capturing performance accurately and reliably across contexts. Therefore, a comparative analysis of these coding schemes is needed to better understand their suitability for specific purposes. This understanding is particularly relevant because the choice of instruments used to assess SSA levels can affect their validity. For example, it is crucial to determine whether a change of perspective—as suggested in one of the coding schemes used in this study—is sufficient to achieve a high level of SSA or whether additional factors need to be fulfilled—as suggested in the other coding scheme used in this study. In addition, such comparative research on coding schemes can contribute to the stated challenge of making learning objectives related to the elaboration of socioscientific issues operational for assessment and teaching (Nielsen, 2020).

Using the assessment triangle (NRC, 2001) as an analytical framework (Shavelson, 2013), the present study explores both potential threats to validity with regard to SSA (i.e. context and coding scheme). To this end, we used a quasi-experimental design in which preservice teachers (biology) engaged in the practice of SSA. Additionally, the observation (i.e. how SSA was measured) and the instruments used for performance interpretation (i.e. how SSA was evaluated) were varied by using four different dilemmas (observation) and two different but similar coding schemes (interpretation). In this way, the influence of these two independent variables (observation and interpretation) on the inferred level of SSA (i.e. dependent variable) was systematically analysed.

The Assessment Triangle

The assessment triangle (Fig. 1) was proposed by the US National Research Council (NRC) in 2001. According to the NRC, the fundamental purpose of educational assessment is to generate data 'that can be used to draw reasonable inferences about what students know' (p. 42). The process of collecting evidence to support certain conclusions is represented as a triad of construct (*What is measured?*), observation (*How is it measured?*), and interpretation (*How can the construct be inferred from the observation?*) in the assessment triangle. For an assessment to be considered effective, it is necessary that each of the three elements be present (NRC, 2001). Hence, the assessment triangle offers a powerful framework for studies aiming to evaluate the validity of test score interpretation in educational research (e.g. Krell et al., 2022). In terms of assessing SSA, the assessment triangle indicates that a clear definition of the research interest (i.e. construct: SSA), a thorough understanding of how the data is collected (i.e. observation: moral-ethical dilemmas), and legitimate inferences based on these data (i.e. interpretation: coding schemes) are necessary.

Construct: What Is Measured?

In the context of the assessment triangle, a construct is a well-defined and literature-grounded concept that represents a specific attribute or ability of interest (NRC, 2001). As a hypothesis for testing, a clearly defined construct enables the differentiation between good and poor performance. To ensure the construct's validity, it is essential to establish a precise definition that is firmly rooted in the existing

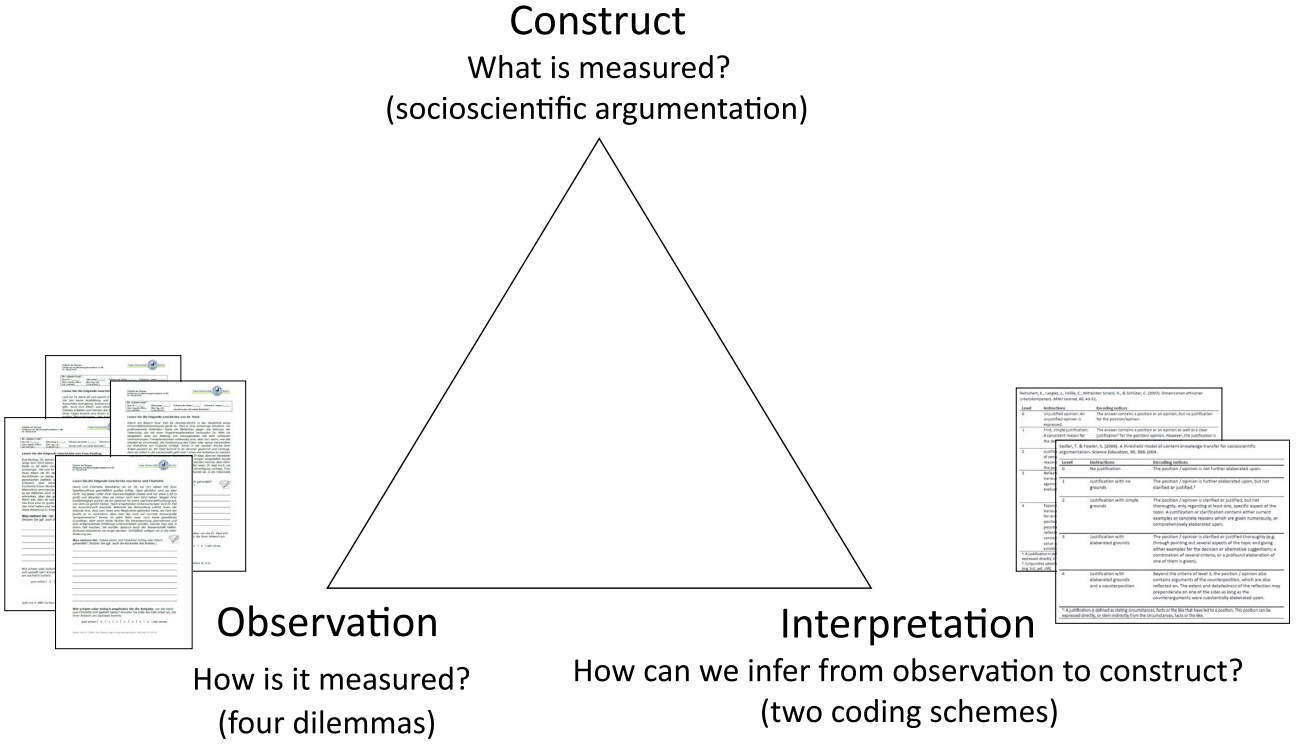

Fig. 1 The assessment triangle, specified for the present study

literature and can effectively differentiate between varying levels of performance (e.g. Messick, 1989; NRC, 2001). The importance of a clear construct definition is also highlighted in the Standards for Educational and Psychological Testing (AERA et al., 2014), which emphasise that 'important validity evidence can be obtained from an analysis of the relationship between the content of a test and the construct it is intended to measure' (p. 14). By demanding a clearly defined construct and grounding it in literature, the assessment triangle supports educators and researchers to develop targeted assessments and make valid inferences about the attribute or ability under investigation.

In science education and in the present study, the construct of SSA is defined as argumentative engagement with socioscientific issues that requires—among others—the evaluation of evidence, the elaboration of multiple and/or conflicting positions, and the inclusion of values and societal aspects (e.g. Christenson & Walan, 2022; Sadler & Donelly, 2006). In addition, a few studies have concluded that SSA performance can be context dependent (e.g. Ladachart & Ladachart, 2021; Sadler & Zeidler, 2005a, 2005b); Topçu et al., 2010) and that it usually comprises several levels that differ in their complexity, which can be used to differentiate between good and poor performance (e.g. Krell et al., 2024; Reitschert et al., 2007; Sadler & Fowler, 2006). In the present study, we refer to the structural complexity of SSA, which is defined as the interplay of multiple arguments that are considered in the argumentation process (see Krell et al., 2024). Increased structural complexity is thus characterised by a greater interplay of arguments.

Observation: How Is It Measured?

Observation describes the performance of an individual on a single task or on a series of tasks, where the 'universe of possible tasks and responses for observing performance, […] logically follows from the definition of the construct' (Shavelson, 2013, p. 78). As discussed by Mathesius and Krell (2019), and regarding the assessment of SSA, there is a huge number of possible tasks that include, for example, different contexts and test formats (e.g. performance assessment, open-ended or closed tasks), which can impact the cognitive demands of a task, which in turn affect the observed performance (e.g. Martinez, 1999; Sadler & Zeidler, 2005a). In this study, we focus on different contexts as variation of the observation, and not, for example, the type of test format.

The term *context* is used in various ways in science education (Gilbert, 2006). In this paper, the term context is understood as proposed by Krell et al. (2015), who equate it with 'task context', that is the task stem or an item characteristic in the assessment instrument itself. Finkelstein (2005, p. 1192) further defines the task context as 'the storyline of a problem' which is given in the task.

Many studies have examined SSA in only one context or context area (e.g. Cetin et al., 2014; Garrecht et al. 2021, Krell et al. 2024), so it remains unclear whether the inferred level of SSA is independent of the respective context (i.e. the specific socioscientific issue) and, hence, generalisable (Shavelson, 2013). Studies investigating SSA in different contexts obtained inconsistent results: Some showed that the context can influence SSA due to relevant content knowledge (e.g.Baytelman et al., 2020; Sadler & Zeidler, 2005a) or affective features (e.g. motivation; see Zeidler, 2014). However, other studies found no differences between SSA levels (i.e. quality of arguments) in different contexts (Werner et al., 2015; Yalman, 2023) or for participants with differing depths of content knowledge (Cetin et al., 2014). Hence, in studies that explicitly examined the SSA in different contexts, the results are heterogeneous. In the study by Topçu et al. (2010), for example, which investigated the quality of preservice science teachers' SSA in three context areas (gene therapy, human cloning, and global warming), the quality of SSA was quite similar across contexts on sample level. However, a higher degree of variability was found at the individual level, suggesting some sort of context dependency. In the study by Ladachart and Ladachart (2021), participants were asked to elaborate two culturally influenced socioscientific issues (floating vessels on rivers and releasing lanterns in the sky). Similar

to Topçu et al. (2010), the authors concluded that there was some context dependency as the participants tended to approach the two issues differently. In the study conducted by Kolstø (2006), in which students had to argue about the construction of new power lines and the potential increased risk of childhood leukaemia, it was found that students' reasoning ability is influenced by their prior knowledge of the specific socioscientific issue.

Although a precise understanding of the extent to which contexts provoke different responses is essential with regard to the validity of task sampling (Shavelson, 2013), the findings on the context dependency of SSA do not appear to be conclusively clarified. Thus, our study explored the consistency of SSA in different moral-ethical dilemmas with comparable requirements in terms of prior knowledge.

Interpretation: How Can the Construct Be Inferred from the Observation?

Interpretation means the extent to which conclusions about a person's ability (in relation to the construct) can be drawn from the observed performance validly (Shavelson, 2013). As discussed by Mathesius and Krell (2019), the interpretation of test scores, particularly in the context of complex constructs such as SSA, necessitates the generalisation of an individual's ability based on their overall test score. To ensure the validity of this generalisation, it is essential that the tasks selected are representative of 'the entire universe of tasks' appropriate for assessing the target construct' (Shavelson, 2013, p. 79). This representativeness is crucial because it enables the inference that an individual's performance on the sampled tasks reflects their broader ability within the construct domain. Moreover, a well-sampled task set enhances the generalisability of the observation, minimising the impact of context dependency and other potential biases.

For the specific construct of SSA, the quality of argumentation is often measured and represented in levels defined by the application of coding schemes (e.g. Reitschert et al., 2007; Sadler & Fowler, 2006). To analyse the structural complexity of an argument, many researchers apply Toulmin's argumentation pattern (TAP; Toulmin, 1958) and related coding schemes (for a review, see Chinn, 2006). With regard to TAP, each argument is examined at the micro level and in terms of different components (i.e. data, warrant, backing, rebuttal, and claim). However, an argumentation can also be analysed at the macro level by considering the interplay of several arguments, focusing, for example, on the anticipation of consequences and a reflection on the argumentation process (e.g. Reitschert et al., 2007; Sadler & Fowler, 2006). In the present study, we focus on this structural complexity of SSA to explore the interplay of different arguments (e.g. the consideration of an opposing position). To capture this structural complexity, we used two established coding schemes (i.e. Reitschert et al., 2007; Sadler & Fowler, 2006), which have different requirements for reaching the respective levels, but share several key similarities: Both schemes analyse argumentation at the macro level and consider a greater number and interaction of arguments as an indication of increased structural complexity. However, they differ in their emphasis on counter positions and the prerequisites for reaching the highest level of complexity, with Reitschert et al. (2007) placing more emphasis on changing perspectives and expanding contexts. For a more detailed description of the similarities and differences between the two coding schemes, please refer to the methods section and Table 2 (and see also the online resource for the full coding schemes). The two coding schemes are therefore quite similar, especially in contrast to other instruments such as TAP (which refers to the micro level; Toulmin, 1958) or the SEE-SEP model (which refers to the content complexity of SSA; Chang Rundgren & Rundgren, 2010). Therefore, these two coding schemes appear useful to investigate whether there are systematic differences in the inferred level of SSA, even if the coding schemes are similar. If such differences are found, this finding would emphasise the importance of paying close attention to the selection of a coding scheme that is appropriate for interpreting the respective construct.

Context of the Study

Part of the data for this study was originally collected for teaching purposes: that is, to initiate a discussion in a seminar on SSA for preservice teachers (biology) in Germany. Further data were collected in additional preservice teacher (biology) seminars.

In Germany, preservice teachers most commonly have to complete a 3-year Bachelor programme (Bachelor of Arts) and a 2-year Master programme (Master of Education) at university, including two subjects of study (i.e. two prospective teaching subjects), followed by an 18-month preparatory service to become certified in-service teachers (Neumann et al., 2017). Notably, both degree programmes are concurrently organised teacher education programmes: that is, with disciplinary and pedagogical studies within the same programme (Zuzovsky & Donitsa-Schmidt, 2017). More precisely, the programmes are organised in three strands (Neumann et al., 2019), including courses in the two subjects (e.g. biology), the respective subject education (e.g. biology education), and learning sciences (e.g. general education).

Students in Germany are expected to develop competencies related to SSA and socioscientific decision-making through secondary science education (Steffen & Hößle, 2014). It can hence be considered important that biology teachers in Germany themselves can engage in SSA to introduce their students to this argumentative engagement with socioscientific issues. Therefore, in the common teacher education standards for biology teacher training, preservice teachers (biology) are expected to be able to grasp biology-related issues in different contexts, to evaluate them factually and ethically, and to justify the individual and social relevance of biological topics (i.e. socioscientific decision-making; *Bewertungskompetenz*, Sekretariat der Kultusministerkonferenz, 2008).

The study participants were preservice teachers (biology) enrolled in a teacher training programme for secondary education at two public universities in Germany (Master of Education in biology). Some of the seminars in which data were collected aimed to introduce various methods to engage secondary students in SSA to the preservice teachers (biology). It is important to note that the data collection occurred prior to the implementation of these methodologies and prior to a comprehensive introduction to the subject matter. Therefore, it is unlikely that the seminar content exerted any influence on the students' SSA. One of the methods employed to engage students in SSA was Lind's (2019) approach to foster moral competence, which was also used to facilitate the acquisition of the students' SSA. Key to this approach is the discussion of a moral-ethical dilemma in the classroom, which is perceived as a dilemma by the students (and not only by the teacher) and which splits the class into two groups with opposing views of almost equal size. Therefore, the participants were asked to read a short (about half a page) moral-ethical dilemma, which described a specific (fictitious) person and their decision in a dilemma situation. Participants were then asked to argue in a written open-response format and on paper whether, in their opinion, the person had acted correctly or incorrectly in the presented dilemma (for a more detailed description of the procedure, see chapter 2).

Aims of the Study and Research Questions

Many of the previous studies focused on SSA in a specific context (area) and therefore did not examine whether SSA is a context-dependent construct (for existing studies, see the literature review above). In addition, the question of whether SSA ratings depend on the coding scheme remains open, as different coding schemes focusing on (slightly) different aspects of SSA can be used. Therefore, in the present study, we aimed to explore (1) whether the quality of SSA is independent of which context is used (i.e. observation: moral-ethical dilemma) and (2) whether different results are obtained depending on how SSA is analyzed (i.e. interpretation: coding scheme).

To achieve these aims, we conducted an exploratory analysis of the data gathered from different M. Ed. in Biology seminar contexts. Hence, the following research questions (RQ) were addressed in this study based on a sample of preservice teachers for biology:

RQ1: To what extent do levels of SSA differ across different observations (i.e. moral-ethical dilemmas)?
RQ2: To what extent do the different emphases and requirements of the coding schemes influence the interpretation of SSA levels?

Material and Methods

Methodically, this study employed a quasi-experimental design in which the observation (i.e. by using four different dilemmas) and the instruments used for performance interpretation (i.e. by using two different coding schemes) were systematically varied in a 4×2 design. In detail, all participants read short moral-ethical dilemmas (each about half a page) describing fictitious persons and their decisions in a dilemma situation. The participants then were asked to provide written responses indicating whether, from their perspective, the fictitious person's decisions were correct or incorrect.

As some of the seminars in which data were collected aimed to introduce methods to engage secondary students in SSA, assessment of SSA for this study was not the first priority. Therefore, due to the limited duration of regular seminar sessions, not all participants could complete all four dilemmas. Instead, the number of dilemmas addressed varied depending on the structure of each seminar and the time available for assessment, with some sessions including only two or three dilemmas. To maintain variability and reduce potential bias, different combinations of dilemmas were systematically assigned across seminars. For instance, in seminars where only two dilemmas were used, participants dealt with different pairs of dilemmas instead of responding to the same dilemmas. While this limitation in study design resulted in incomplete randomisation, the careful variation of dilemma combinations allowed that all participants contributed to a diverse set of scenarios. The responses were then analyzed using Excel for coding and SPSS for further analysis.

Participants

To answer our research questions, we analysed responses from preservice teachers (biology) who participated in Master of Education biology seminars and agreed to participate voluntarily and anonymously. Hence, the participants were in the final phase of their teacher education programme at university and had already completed a Bachelor's programme as well as parts of the Master's program. Demographic information beyond enrollment in the biology teacher education programme was not collected because the study design focused exclusively on participants' responses to the dilemmas and we aimed to keep collection of personal data minimal. Data were collected in 2018 ($n = 18$), 2019 ($n = 22$), and 2022 ($n = 29$).

Based on the procedure described above, the dataset comprises a number of participants who worked on two dilemmas ($n = 45$), a small group who worked on three dilemmas ($n = 6$), all in various combinations, and a third group who worked on all four dilemmas ($n = 13$). Five participants worked on just one dilemma (e.g. because they missed a seminar day), so they were not included in the following analysis. As a result, we analysed 160 responses from 64 preservice teachers (biology).

Moral-Ethical Dilemmas

In this study, we used four moral-ethical dilemmas which were originally developed by and taken from Lind (2006, 2019). Each dilemma is fictitious but authentic and describes a problematic situation in a science-related but rather medical context, whereas no specific content knowledge was required to understand the underlying issue. Each dilemma also contains the decision of fictitious persons. The participants had to argue and explain whether they agreed or disagreed with the person's decision in each dilemma. The four moral-ethical dilemmas used in this study referred to the following: (1) extrauterine fertilisation with experimental DNA modification (=childbearing), (2) abortion of a potentially disabled child (=abortion), (3) organ transplantation without consent (=transplantation), and (4) sale of embryos (=embryo selling). The four dilemmas differ in the extent to which they explicitly address numerous aspects that may influence the participants' SSA, such as the consequences of the decision for the individual or others, and the explicit reference to religious beliefs (Table 1).

Coding

In the present study, we categorised participants' SSA on four different moral-ethical dilemmas according to their structural complexity. Their SSA was analyzed using two different coding schemes in each case (i.e. Reitschert et al., 2007; Sadler & Fowler, 2006).

Both coding schemes distinguish between five levels of structural complexity (0–4), but each scheme has different emphases, and the statements are scored slightly differently (Table 2; see online resource for the full coding schemes). Common to both schemes is that as the level increases, the complexity of the SSA increases (e.g. by weighing up the arguments, giving examples or detailed reasons). However, there is a clear difference in the extent to which the inclusion of the respective counter positions is considered. In Sadler and Fowler's (2006) scheme, this is merely relevant for reaching the highest level (level 4; Justification with […] a counter position). In Reitschert et al.'s (2007) scheme, it is already a prerequisite for level 2 (Justification with a change of perspective). A further difference between the two coding schemes lies in the prerequisites for reaching the highest level. Whereas Sadler and Fowler's (2006) scheme requires not only detailed reasoning but also consideration of a counter position, Reitschert et al.'s (2007) scheme demands an expansion of the context to other epochs or value systems.

Our qualitative data analysis considered a range of quality-ensuring procedures and followed the approach of qualitative content analysis (Göhner & Krell, 2020); Schreier, 2012). First, two well-established coding schemes were used. Second, five responses to each moral-ethical dilemma were coded by a research assistant using the two coding schemes. As part of this process, the coding schemes were adapted to the collected data by sharpening the level description and adding sample responses. For example, more precise definitions were provided for what constitutes an 'expansion of the context' and what is meant by 'connection to other value systems' and illustrative sample responses from participants were included for each level. The adjustments were then reviewed by the first and last authors of this paper and revised as necessary to ensure clarity and consistency. Subsequently, these two authors coded the same five responses to each moral-ethical dilemma and discussed them together with the research assistant. Thereafter, all remaining responses were coded independently by two coders (the research assistant and one of the authors) using the revised coding schemes to determine intercoder agreement as a measure of coding objectivity (Göhner & Krell, 2020). The research assistant also coded all responses to each moral-ethical dilemma twice using both coding schemes, with at least 1 week between the first and the second coding. On this basis, intracoder agreement was calculated as a measure of coding reliability (Göhner & Krell, 2020). Cohen's Kappa (Table 3), calculated as suggested by Brennan and Prediger (1981), indicated a substantial to almost perfect intracoder agreement ($0.79 \leq K \leq 1.00$) and a

Table 1 Overview of the aspects that are explicitly (i.e. literally) addressed in the four dilemmas at the level of the decision-maker and the decision

	Childbearing	Abortion	Transplantation	Embryo selling
Decision-maker: Personal context	Both suffer[NOTE] from their short stature and from the fact that they still don't have a child; great business success with their company	Renounced having a child for a long time due to her successful career; would like to have a child before too old	Is just started out as assistant physician	Lives in a poor, South American country; no formal training and no job; parents also without work; younger siblings must work and cannot attend school
Religious beliefs	No information about religious beliefs	No information about religious beliefs	Practising Catholic	Raised strictly according to Catholic principles
Decision: Potential consequences (for decision-maker and others)	Decision-maker: Must take responsibility for any difficulties that may arise as consequence of the experimental DNA modification Others: No possible consequences for others mentioned	Unclear: Depending on the sex of the unborn child, either severe disability or carrier of the genetic defect Decision-maker: Prepared to carry the responsibility and resources Others: Decision contradicts the man's wishes; the unborn is aborted	Decision-maker: No possible consequences were mentioned Others: Disturbance of the repose; not enough skin grafts a for a seriously injured patient	Decision-Maker: Receives training as a teacher, receives good medical care, must sell embryos once a year for 5 years Others: Enough money to feed the family
Legality	Unclear: No legal regulation yet	No information, but in the country of data collection, abortion remains unpunished for up to the 12th week of pregnancy	Illegal: Consent is legally mandated	No information

[Note]The four dilemmas were chosen as they represent an established instrument. The presentation of the content does not reflect the authors' view or choice of words

Table 2 Descriptions of the two coding schemes (see online resource for the full coding schemes)

Level	Reitschert et al. (2007)	Sadler and Fowler (2006)
0	Opinion: An opinion is expressed without further explanation	No justification
1	Simple justification: A consistent reason for the decision is provided	Justification with no grounds
2	Justification with a change of perspective: Various reasons for and against the position are stated	Justification with simple grounds
3	Reflected justification: Various reasons for and against the position are evaluated	Justification with elaborated grounds
4	Expansion of context: Various reasons arguing for and against the position are evaluated, possible consequences are reflected upon, and a connection to e.g., other value systems or epochs is established	Justification with elaborated grounds and a counter position

Table 3 Intercoder agreement and intracoder agreement: Kappa values, calculated as suggested by Brennan and Prediger (1981)

	Intercoder agreement		Intracoder agreement	
	Sadler and Fowler (2006)	Reitschert et al. (2007)	Sadler and Fowler (2006)	Reitschert et al. (2007)
Childbearing	0.62	0.46	0.82	0.82
Abortion	0.57	0.60	1.00	1.00
Transplantation	0.95	0.82	1.00	1.00
Embryo selling	0.72	0.57	0.79	0.79

moderate to almost perfect intercoder agreement ($0.46 \leq K \leq 0.95$) (Landis & Koch, 1977). Finally, in order to achieve a consensus for the data analysis, all deviating codings were resolved by discussion.

Data Analysis

To address RQ1, we compared participants' levels of SSA between the four moral-ethical dilemmas. To address RQ2, we compared the levels of SSA between the two coding schemes. In both cases, differences (between the dilemmas or the coding schemes) would indicate the need to carefully consider the respective approach used for the observation (i.e. moral-ethical dilemma) and the interpretation (i.e. coding scheme) of SSA in order to reach alignment between the triad of construct, observation, and interpretation (NRC, 2001).

As not all participants worked on all moral-ethical dilemmas and additionally on different combinations of the dilemmas, we analysed the data both in a between-subjects comparison (univariate ANOVA) and in a within-subjects comparison (Wilcoxon tests). For the former, we organised the data independently of the participant. For the analysis, we tested for variance homogeneity of the samples (Levene test; fulfilled) and used two-factorial ANOVAs. Post hoc Gabriel tests were conducted with the independent variables 'dilemma' (childbearing vs. abortion vs. transplantation vs. embryo sale) and 'coding scheme' (Sadler and Fowler vs. Reitschert et al.), as well as the dependent variable 'SSA level'. In order to investigate the question of whether someone who achieves a high score in one dilemma also achieves a high score in another dilemma, regardless of whether the person has worked on all dilemmas or the same combination of dilemmas, a within-subjects analysis was carried out with all pairwise dilemma combinations. This procedure also allowed to determine which dilemmas were associated with comparably high scores and which differed in this respect.

Results

Descriptive Analysis

The participants' SSA revealed high variance, ranging from the lowest (level 0) to the highest level (level 4) in both coding schemes (Table 4, Table 5). The mean scores (M) represent the average SSA level achieved by participants for each dilemma, providing an overall indication of the complexity and elaboration of their argumentation. Higher mean scores indicate that participants, on average, reached higher SSA levels and demonstrated more structurally complex argumentation. On average, the mean SSA levels ranged from $M = 1.58$ (abortion) to $M = 2.93$ (embryo selling) for both coding schemes. When coded according to Sadler and Fowler's coding scheme, participants achieved the highest level on the dilemma about the sale of embryos by a poor young woman (embryo selling; $M = 2.93$) and the second highest level on the dilemma about illegal organ transplantation (transplantation; $M = 2.89$).

Table 4 Number of responses coded in each level for both coding schemes

Sadler and Fowler (2006)								
	N_{L0}	N_{L1}	N_{L2}	N_{L3}	N_{L4}	N	M	SD
Childbearing	1	5	13	11	**16**	46	2.78	1.11
Abortion	1	5	**15**	13	6	40	2.45	0.99
Transplantation	0	0	**11**	9	8	28	2.89	0.83
Embryo selling	0	5	12	10	**19**	46	2.93	1.06
Reitschert et al. (2007)								
	N_{L0}	N_{L1}	N_{L2}	N_{L3}	N_{L4}	N	M	SD
Childbearing	5	**18**	7	4	12	46	2.00	1.41
Abortion	5	**16**	13	3	3	40	1.58	1.06
Transplantation	0	**16**	3	5	4	28	1.89	1.17
Embryo selling	2	**13**	9	9	**13**	46	2.39	1.29

The levels that the majority of participants have reached in the respective dilemma are highlighted in bold

L, level; *M*, mean score representing the average SSA level achieved by participants; *SD*, standard deviation

Table 5 Sample quotes for each level and the two coding schemes

Level	Reitschert et al. (2007)	Sadler and Fowler (2006)
0	There is no right or wrong. This decision has to be made subjectively on the basis of many criteria. If she feels ready, she should do it. From the child's point of view, or with the child in mind, I would perhaps argue differently. (0203DAHA)	No statement was coded as 0
1	Wrong, there are clear regulations for this, and every person should be able to determine his or her body even after death. (6012GARO)	I cannot make any statement about it, because I have neither information about the seriousness of the company, nor do I have insight into the contract. An evaluation with my current state of knowledge would be premature. (0224KIKI)
2	In my opinion, Lara acted correctly. She weighed different aspects (family, faith, future) and decided to put the emphasis on her family and her future. (2730REMA)	In my opinion, she acted correctly. I find it incomprehensible to abort one's child because of a possibly genetic defect/disability. (2205CAHA)
3	I can understand the decision because of the situation and I think it is right. Even though she is acting against her faith, she can perhaps comfort herself with the thought of contributing to saving many other lives. (1014CLJO)	They have acted correctly because they themselves know life as a small person. This seems to bring many problems with it. Since they themselves conclude that they are doing their child a great favor, I think their decision is right. (4326HERA)
4	Right, because the improved life circumstances outweigh their moral worries and 'heavy thoughts'. However, only because of their precarious plight. Feeding her family is also more important to her than genetic cures. (0102ANJÖ)	I am really a proponent of organ donation and think everyone should donate after death to save lives. However, there was no consent in this case. The doctor acted against her 'codex'. The decision was wrong. (117ROHE)

When coded according to Reitschert et al.'s scheme, results showed an approximately similar pattern: Participants also reached the highest level on the embryo selling dilemma (embryo selling; $M = 2.39$). Regardless of the coding scheme, participants scored lowest on the dilemma of aborting a potentially disabled child (abortion; Sadler and Fowler: $M = 2.45$; Reitschert et al.: $M = 1.58$). In addition, when analysed according to the coding scheme of Sadler and Fowler, only very few answers were assigned to the lowest level ($n = 2$), but many to the highest level ($n = 49$). When analysed according to the scheme of Reitschert et al., the two extremes were somewhat more balanced. Here, 12 of the responses were assigned to the lowest level and 31 to the highest level (Table 4).

Table 5 provides sample quotes for each level and the two coding schemes. Please note that the full coding schemes (including the sample quotes) can be found in the online resource of this article.

Levels of SSA Across Different Moral-Ethical Dilemmas (RQ1)

The between-subjects comparison (univariate ANOVA) revealed no differences in the SSA levels between the four moral-ethical dilemmas when analyzed according to Sadler and Fowler's coding scheme ($F_{[3, 156]} = 1.745$, $p = 0.160$, $\eta^2 = 0.032$, $n = 160$). In contrast, when coding according to Reitschert et al., a significant difference was found ($F_{[3, 156]} = 3.058$, $p = 0.030$, $\eta^2 = 0.056$, $n = 160$). However, post hoc Gabriel tests revealed significant differences in the SSA levels only between the two dilemmas of abortion and selling embryos ($p = 0.019$; Fig. 2).

The within-subjects comparison (Wilcoxon test) revealed significant differences between the dilemmas in only four of the 12 pairwise comparisons (Table 6). Two of these pairwise comparisons relate to the comparison between the abortion and embryo selling dilemma, which, according to Reitschert et al.'s coding scheme, yielded also significantly different levels in the between-subjects comparison. For example, participant 0910ANRE wrote the following in the abortion dilemma:

> Correct: It is always possible for a child to be born disabled or to suffer a disability later on due to other circumstances. An abortion (unintentional!) can demonstrably lead to depression in the woman and in general it would be judged here that a person who is disabled from birth is not worth living.

This response was coded as level 1 based on the Reitschert et al. coding scheme because it includes a consistent reason for the decision but not a change of perspective. In contrast, the same participant wrote the following in the embryo selling dilemma:

> From a purely material point of view, it seems to be the right decision, as everyone involved gains an advantage. Morally speaking, it is the wrong decision. The verdict depends on which value is more important to Lara. I tend to lean towards wrong, as the pharmaceutical company is clearly taking advantage of Lara's predicament. This would not be possible in a western industrialized country.

This response was coded as level 4 based on the Reitschert et al. coding scheme because it includes a reflected justification (i.e. reasons for and against the position are evaluated) and also refers to the value system of different countries.

Levels of SSA Depending on the Coding Scheme (RQ2)

The analysis revealed a consistent difference between the two coding schemes, regardless of the dilemma presented. Participants achieved significantly higher levels of SSA in the Sadler and Fowler coding scheme compared to the Reitschert et al. scheme ($F_{[1, 63]} = 107.74$, $p < 0.001$, $\eta^2 = 0.73$, $n = 64$; Fig. 3).

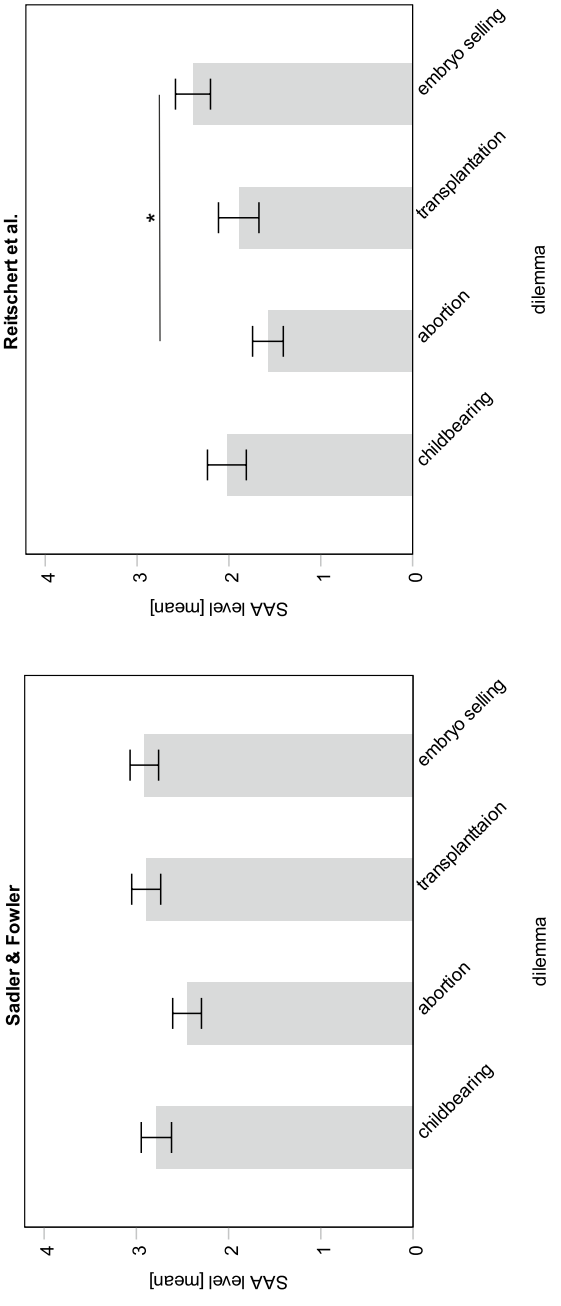

Fig. 2 Levels achieved in the four moral-ethical dilemmas (mean ± SE). $n = 160$; $^{*}p < 0.05$

Table 6 Pairwise individual comparisons (Wilcoxon test) between two moral-ethical dilemmas for each coding scheme

		Sadler and Fowler (2006)					Reitschert et al. (2007)				
	n	*Z*	*p*	*r*	*M*	*SD*	*Z*	*p*	*r*	*M*	*SD*
Childbearing —abortion	31	−0.69	0.49	0.12	0.16	1.27	−1.43	0.15	0.26	0.42	1.59
Childbearing —transplantation	18	−1.47	0.14	0.35	−0.50	1.30	−0.18	0.86	0.04	−0.06	1.55
Childbearing —embryo selling	28	−1.35	0.18	0.26	−0.32	1.25	−2.18	**0.03**	0.41	−0.71	1.61
Abortion — transplantation	16	−2.13	**0.03**	0.53	−0.63	1.09	−0.98	0.33	0.25	−0.38	1.36
Abortion — embryo selling	23	−2.33	**0.03**	0.49	−0.70	1.43	−2.77	**0.01**	0.58	−1.22	1.76
Transplantation — embryo selling	25	−0.42	0.68	0.08	0.16	1.28	−0.80	0.42	0.26	−0.28	1.65

Significant differences at the $p < 0.05$ level are highlighted in bold

M, mean score; *SD*, standard deviation

Furthermore, there was no significant interaction effect between the coding scheme and the dilemma ($F_{[3, 312]} = 0.602$, $p = 0.614$, $\eta^2 = 0.006$, $n = 320$; Fig. 4), indicating that the difference between the two coding schemes is independent from the specific dilemma.

Discussion

As summarised in the literature review, many of the existing studies have focused on assessing SSA in one particular context or context area (e.g. Garrecht et al., 2021) and analysed the performance using one particular coding scheme (e.g.Cetin et al., 2014; Christenson et al., 2014). Thus, most of the previous studies have not systematically examined whether SSA performance is independent of how SSA is measured (i.e. observation) or how the construct can be inferred from the observation (i.e. interpretation). In the present study, both potential threats to the valid assessment of SSA were explored in a sample of preservice teachers (biology).

Observation: Levels of SSA Across Different Moral-Ethical Dilemmas

To address RQ1, that is to systematically examine to what extent the participants' levels of SSA differ across different observations (i.e. moral-ethical dilemmas), we analysed their SSA on four different moral-ethical dilemmas, which differ in two essential points: firstly, in terms of the specific issues at stake (childbearing, abortion, transplantation, embryo selling); and secondly, in terms of the extent to which various factors that may influence SSA are explicitly addressed. These factors include, for example, the personal background or religious beliefs of the decision-maker, as well as the possible consequences of the decision and its legality (Table 1).

All participants included in the data analysis dealt with at least two of the four moral-ethical dilemmas. The results of our study show that the levels of SSA remained relatively stable across most dilemmas (i.e. the only significant difference was detected between the dilemma of abortion and the dilemma of selling embryos when using the coding scheme by Reitschert et al.). Similar to the study by Topçu et al. (2010), there were more differences at the individual level (within-subjects comparison) than at the sample level (between-subjects comparison). For the significant differences that were found at the individual level (four out of 12 possible combinations were found to be significant, i.e. one-third), the effect sizes were rather small to medium. However, for most comparisons (two-thirds), the participants reached a similar level of SSA. In other words, someone who has reached a high level in one dilemma is likely to also reach a high level in another dilemma. This confirms

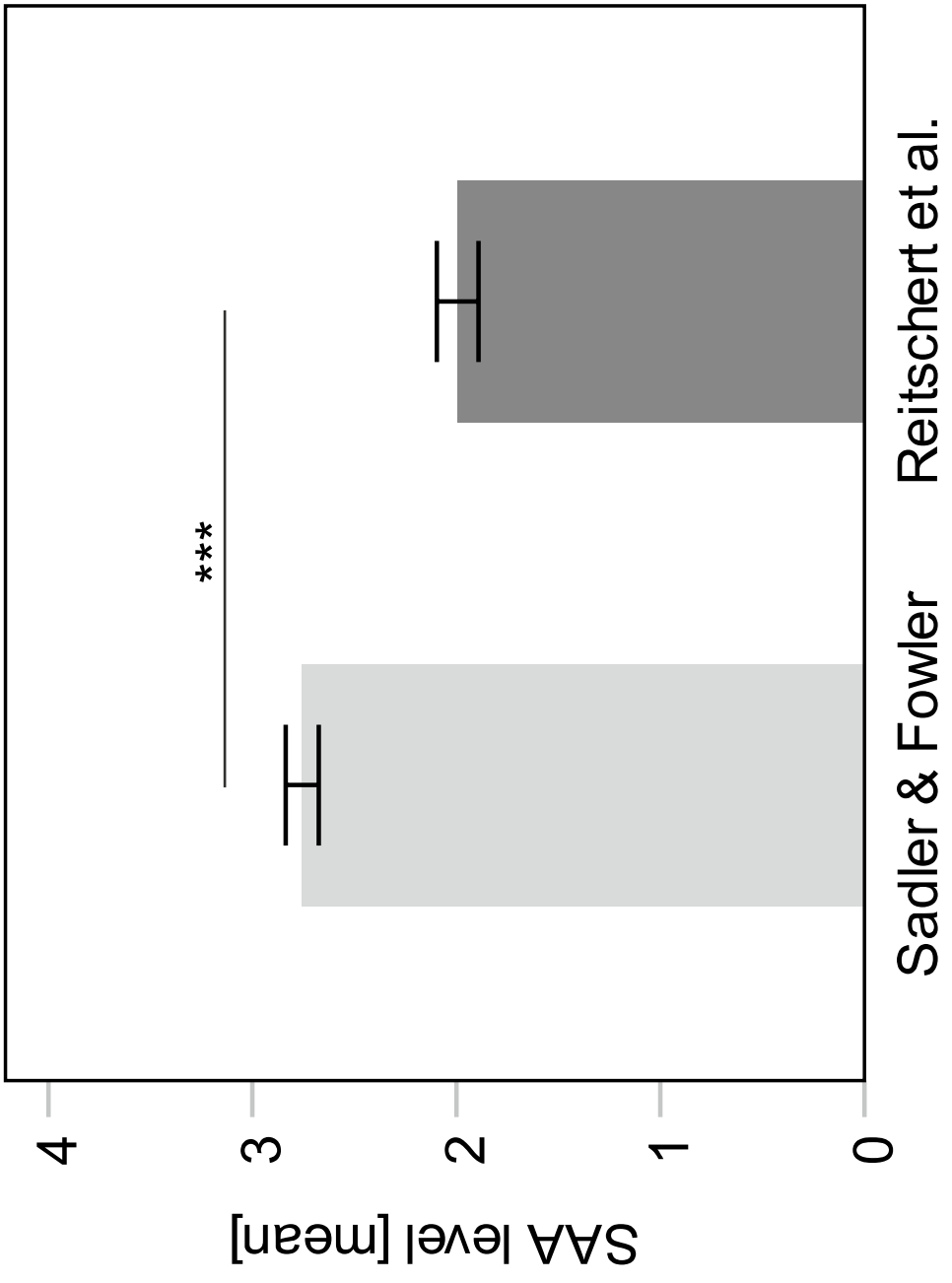

Fig. 3 Levels achieved in the two coding schemes across all dilemmas (mean ± SE). $n = 64$; $^{***}p < 0.001$

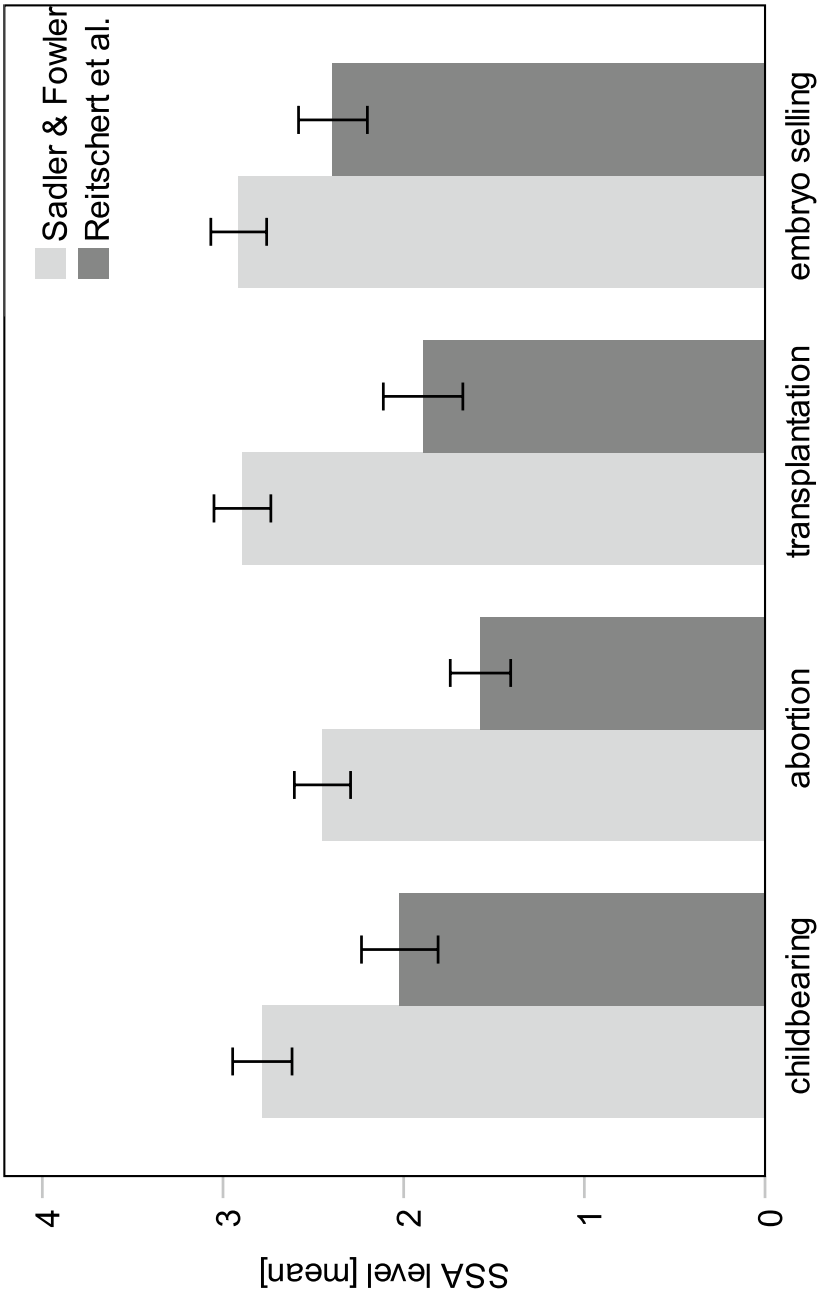

Fig. 4 Levels achieved in the two coding schemes in the four moral-ethical dilemmas (mean ± SE). $n_{\text{childbearing}} = 46$, $n_{\text{abortion}} = 40$, $n_{\text{transplantation}} = 28$, $n_{\text{embryo selling}} = 46$

earlier findings that the quality of SSA seems relatively independent of context (Werner et al., 2015; Yalman, 2023) and supports the notion that SSA appears to be a somewhat stable trait across various contexts. This stability has important implications for educational practices, as it facilitates the consistent development and assessment of argumentation skills across different subjects and scenarios. For assessments, it hints toward the reliability and validity of evaluating SSA in diverse contexts. This suggests that the specific choice of dilemmas might be less critical, provided that the dilemmas are comparable in terms of the type of information presented.

In contrast to our findings, other studies have found some influence of context in terms of the required content knowledge (e.g. Baytelman et al., 2020) or the context-specific motivation (e.g. Topcu et al., 2010). We cannot rule out the possibility that content knowledge also has an influence in our study, but the four moral-ethical dilemmas that we used did not require extensive content knowledge, and if they did, the information required to assess the situation was provided in the dilemma (e.g. the effects of the possible disability in the abortion dilemma). However, context-specific motivation may have played a role in the participants' SSA, as the dilemmas vary in terms of the extent to which various factors that could influence motivation are addressed (e.g. religious beliefs; Table 1). These differences with regard to the explicit naming of various factors may also influence the extent to which they are included in the argumentation and thus also the level of SSA to which individuals are assigned. For example, religious beliefs are explicitly mentioned in only two of the dilemmas (transplantation and embryo selling), and only in one dilemma is it *explicitly* stated that the individual's actions are illegal (transplantation). Therefore, it seems more likely that these aspects will be addressed in the corresponding SSA, and participants with similar or different religious beliefs may be particularly motivated. This argumentation is supported by the observation that the dilemma about the woman selling her embryos reached the highest level in both coding schemes, which could be related to the specific content of the dilemma: In contrast to the other dilemmas, it involved a life-threatening situation. In addition, this dilemma encourages consideration of other value systems and religious beliefs, as it takes place in a South American country, and it is mentioned that the decision contradicts the woman's religious beliefs. This might favour the consideration of a (counter) position (i.e. the life-threatening situation, conflicting religious beliefs) and the expansion of the context (i.e. by making a connection to other value systems), which in turn are prerequisites for reaching the highest level of SSA (i.e. counter position in Sadler and Fowler and expansion of context in Reitschert et al.). In contrast, the abortion dilemma, which has the lowest level in both coding schemes, revolves around a successful businesswoman who is not in a life-threatening situation and has likely financial means to deal with the uncertain future situation (e.g. hiring care or purchase of supportive devices for the potentially disabled child). Moreover, this dilemma does not explicitly address contradictory beliefs or other value systems. Consequently, the participants are not directly made aware of aspects that would lead to higher levels of coding when they are addressed in their SSA. Nevertheless, as we did not capture the personal situations of the participants or their religious beliefs, we can only speculate on the extent to which the mentioning of specific aspects has influenced their SSA.

In summary, our results indicate that the specific dilemma can but does not necessarily influence the level of SSA (RQ1). Therefore, the findings suggest that the context is not a major threat to a valid assessment of SSA, as long as extensive content knowledge is not required (which may be the case for other socioscientific issues) and the dilemmas are largely analogous in that they refer to factors that could potentially impact the SSA (e.g. religious beliefs and background of the decision-maker, potential consequences of the decision, and its legality). Researchers should therefore carefully consider the extent to which the respective dilemma contains prerequisites that must be fulfilled in order to achieve a high level in the coding scheme used. Furthermore, we presented participants four moral-ethical dilemmas in which they had to judge the decision of another person, but other more open scenarios could also be used for the assessment of SSA (e.g. controversial societal issues such as a mandatory vaccination (Krell et al., 2024)) which might lead to different results regarding context dependency.

Interpretation: Levels of SSA Depending on the Coding Scheme

To address RQ2, that is to investigate the extent to which the different emphases and requirements of the coding schemes influence the interpretation of SSA levels, we analysed all responses using two different coding schemes. Although both coding schemes focus on the structural complexity of SSA at the macro level and, thus, are relatively similar, we found significant differences between the levels of SSA that the participants achieved with each scheme: Participants achieved significantly higher levels of SSA (large effect size; $\eta^2 = 0.73$) when coded according to the Sadler and Fowler scheme compared to the levels reached when coded according to the Reitschert et al. scheme. In addition to slight differences in the weighting of arguments and the inclusion of an expanded context, this difference is probably mainly due to the fact that the inclusion of a counter position is weighted differently in the two coding schemes: In Reitschert et al.'s scheme, the consideration of a counter position is already a prerequisite for level 2, whereas in Sadler and Fowler's scheme it is only relevant for reaching the highest level (i.e. level 4). Because of these differences, the rating of SSA levels varies depending on the coding scheme used, to the extent that a SSA is assigned to level 4 in Sadler and Fowler (2006) because it contains a counter position, while in Reitschert et al. (2007) it is assigned to level 2 due to the lack of an evaluation of various reasons for and against the chosen position.

Summarising, the choice of coding scheme seems to have a major influence on the interpretation (RQ2). This strong dependence of the assigned level on the coding scheme emphasises how careful the selection of instruments for interpreting SSA should be. Therefore, one should consider thoroughly which coding scheme to use depending on the research question (for researchers) or the goals of improving SSA among individuals (for educators). As mentioned above, this becomes especially important in connection with the chosen dilemma because the dilemma itself may contain factors that, to varying degrees, include certain conditions that must be met in order to achieve a high level in the respective coding scheme, even if we could not prove this in our study. Thus, with regard to the two main differences between the coding schemes of Reitschert et al. and Sadler and Fowler, one should ask the following question: How important is the consideration of the counter position and the extension of the context (e.g. to other value systems)? More generally, as our findings suggest strong differences in the level of SSA depending on the coding scheme, one can ask which aspects of SSA are crucially important in globalised societies and to what extent established coding schemes cover these aspects. From this perspective, the inclusion of context extension appears to be a valuable and relevant addition.

In the present study, we found evidence that the coding scheme used for interpretation and, to a certain exten56 also the context have an influence on the level of SSA. Our findings further suggest that the way in which the construct is inferred from the observation (interpretation) threatens the validity of the assessment of SSA more than the way in which SSA is measured (observation). Validity has been defined as 'the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests' (AERA et al., 2014, p. 11). Hence, validity always depends on the intended uses of the tests. Therefore, careful identification and use of coding schemes that match the proposed use of an assessment are a critical aspect of assessing SSA.

Limitations and Future Studies

With respect to the present study, some limitations should be noted. First, the use of other topics (e.g. with more personal relevance or more media coverage) could lead to different results in participants' SSA. However, by using four different moral-ethical dilemmas, we have attempted to address this limitation by providing participants with a broad variety of contexts. Secondly, with regard to the sample, it is critical to note that this study analysed data from preservice teachers and only a handful of participants answered all four dilemmas due to organisational circumstances. This unequal distribution among the

four dilemma combinations may influence the statistical power as well as the generalisability of the results. If all participants had answered all four dilemmas, a broader in-between-subjects comparison could have been made which would have provided more detailed information on the extent to which a person's SSA differs between observations and depends on the coding scheme. Hence, future studies can build on the findings of the present exploratory study by varying dilemmas systematically, and by collecting data from populations other than preservice teachers. Thirdly, we assessed SSA with only two coding schemes. Using other coding schemes would probably lead to even greater differences regarding the assigned SSA levels. However, we deliberately chose two schemes that are relatively similar as they both refer to macro-level structural complexity. Future studies should continue to explore SSA in other contexts and settings, such as contexts that might be more relevant to the lives of teachers (e.g. classroom situations). In addition, potential influencing factors (e.g. argumentation ability, individual valence, knowledge of the context) should be systematically recorded and/or varied.

Implications

This study contributes to the research on the systematic assessment of SSA. In particular, our comparative analysis of contexts and coding schemes can assist in addressing the challenge of operationalising socioscientific issue-related learning objectives for assessment and teaching, as outlined by Nielsen (2020). Researchers and practitioners in science education aiming to assess or improve students' SSA should carefully consider which coding scheme aligns best with their goals and the specific curriculum, as our findings revealed significant variance in the assigned levels, depending on the chosen scheme (or more specifically, which aspect of SSA is emphasised in the respective scheme), despite the similarity of both coding schemes in evaluating structural complexity at the macro level. Because of these differences, researchers in science education should also be very careful in selecting the most appropriate coding scheme for their research question, not only in terms of whether SSA is evaluated with regard to structural complexity at the micro level (i.e. TAP; Toulmin, 1958) or at the macro level (e.g. Reitschert et al., 2007; Sadler & Fowler, 2006), or in terms of its content complexity (SEE-SEP; Chang Rundgren & Rundgren, 2010), but also with regard to which scheme is used within each level.

Our findings further suggest several practical implications for classroom practice: Because participants' SSA levels were largely consistent across dilemmas of comparable complexity and content, teachers might carefully select ethical dilemmas that are engaging and relevant for students without the specific context substantially affecting the quality of their reasoning, given that they all present the necessary information in a similar manner. Moreover, the observed differences between coding schemes underscore the importance of incorporating counter positions and broader contexts in classroom activities. Teachers can actively encourage students to incorporate alternative perspectives, promoting additional aspects that support higher SSA levels. Finally, since individuals who demonstrate high SSA in one dilemma tend to perform similarly in others, as shown by the pairwise individual comparisons, educators can address weaknesses through targeted support or leverage strengths through structured peer discussions to foster complex argumentation across the class.

However, the results of this study suggest that the moral-ethical dilemmas used to assess the structural complexity of SSA might be interchangeable, as long as the dilemmas do not require extensive content knowledge and are more or less comparable in terms of the extent to which they evoke, for example, the inclusion of a counter position or a different value system.

Author Contribution Conceptualisation: M. Krell. Methodology: N. Minkley, M. Krell. Formal analysis and investigation: N. Minkley, M. Krell. Writing—original draft preparation: N. Minkley. Writing—review and editing: N. Minkley, C. Garrecht, M. Krell. Resources: M. Krell, N. Minkley.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability Data are available on request from the corresponding author.

Declarations

Ethics Approval and Consent to Participate The study was approved by the Ethics Committee of the Professional School of Education of the Ruhr-Universität Bochum (Approval number: EPSE-2022.003) and all participants gave written informed consent.

Conflict of Interest The authors declare no competing interests.

References

AERA, APA, NCME (2014) Standards for educational and psychological testing: National Council on Measurement in Education, American Educational Research Association, Washington DC

Baytelman A, Iordanou K, Constantinou CP (2020) Epistemic beliefs and prior knowledge as predictors of the construction of different types of arguments on socioscientific issues. J Res Sci Tech 57(8):1199-1227. https://doi.org/10.1002/tea.21627

Bencze L, Pouliot C, Pedretti E, Simonneaux L, Simonneaux J, Zeidler D (2020) SAQ, SSI and STSE education: Defending and extending "science-in-context". Cult Stud Sci Educ, 15: 825–851. https://doi.org/10.1007/s11422-019-09962-7

Brennan, RL, Prediger, DJ (1981) Coefficient Kappa. Educ Psychol Meas 41:687–699. https://doi.org/10.1177/001316448104100307

Cetin PS, Dogan N, Kutluca AY (2014) The quality of pre-service science teachers' argumentation. J Sci Teach Educ 25(3):309–331. https://doi.org/10.1007/s10972-014-9378-z

Chang Rundgren SN, Rundgren C (2010) SEE-SEP. From a separate to a holistic view of socioscientific issues. APFSLT 11(1):1–24. https://www.eduhk.hk/apfslt/v11_issue1/changsn/index.htm. Accessed 8 Aug 2025

Chinn C (2006) Learning to argue. In: O'Donnell AM, Hmelo-Silver CE, Erkens G(ed) Collaborative learning, reasoning, and technology. Erlbaum, New York, pp 355-383

Christenson N, Walan S (2022) Developing Pre-service teachers' competence in assessing socioscientific argumentation. J Sci Teach Educ 8(6):1–23. https://doi.org/10.1080/1046560X.2021.2018103

Christenson N, Chang Rundgren SN, Zeidler DL (2014) The relationship of discipline background to upper secondary students' argumentation on socioscientific issues. Res Sci Educ 44(4):581–601. https://doi.org/10.1007/s11165-013-9394-6

DeBoer GE (2000) Scientific literacy: Another look at its historical and contemporary meanings and its relationship to science education reform. J Res Sci Teach 37(6):582–601. https://doi.org/10.1002/1098-2736(200008)37:6<582::AID-TEA5>3.0.CO;2-L

Finkelstein N (2005) Learning physics in context: A study of student learning about electricity and magnetism. Int Journal Sci Educ 27(10): 1187–1209. https://doi.org/10.1080/09500690500069491

Garrecht C, Reiss MJ, Harms U (2021) 'I wouldn't want to be the animal in use nor the patient in need'–the role of issue familiarity in students' socioscientific argumentation. Int J Sci Educ 43(12): 2065–2086. https://doi.org/10.1080/09500693.2021.1950944

Gilbert JK (2006) On the nature of "context" in Chemical Education. Int J Sci Educ 28(9):957-976. https://doi.org/10.1080/09500690600702470

Göhner M, Krell, M (2020) Qualitative Inhaltsanalyse in naturwissenschaftsdidaktischer Forschung unter Berücksichtigung von Gütekriterien: Ein Review. [Qualitative content analysis in science didactic research underconsideration of quality criteria]. ZfDN 26(1):207–225. https://doi.org/10.1007/s40573-020-00111-0

Jiménez-Aleixandre MP, Erduran S (2008) Argumentation in science education: An overview. In: Erduran S, Jiménez-Aleixandre MP (eds), Argumentation in science education. Springer Netherlands, Dordrecht, pp 3-28

Kolstø SD (2006) Patterns in students' argumentation confronted with a risk-focused socio-scientific issue. Int J Sci Educ 28(14):1689–1716. https://doi.org/10.1080/09500690600560878

Krell M, Upmeier zu Belzen A, Krüger D (2014) Context-specificities in students' understanding of models and-modelling: An issue of critical importance for both assessment and teaching. In: Constantinou C, Papadouris N, Hadjigeorgiou A (eds) E-Book proceedings of the ESERA 2013 conference. Science education research for evidence-based teaching and coherence in learning. Part 6. Nature of science: History, philosophy and sociology of science. Science Education Research Association, Nicosia

Krell M, Reinisch B, Krüger D (2015) Analyzing students' understanding of models and modeling referring to the disciplines biology, chemistry, and physics. Res Sci Educ 45(3):367–393. https://doi.org/10.1007/s11165-014-9427-9

Krell M, Xu KM, Re GD, Paas F (2022) Editorial: Recent approaches for assessing cognitive load from a validity perspective. Front Educ 6:838422. https://doi.org/10.3389/feduc.2021.838422

Krell M, Garrecht C, Minkley, N (2024) Preservice biology teachers' socioscientific argumentation: Analyzing structural and content complexity in the context of a mandatory COVID-19 vaccination. Int J Sci Math Educ 22:121–141. https://doi.org/10.1007/s10763-023-10364-z

Ladachart L, Ladachart L. (2021). Preservice biology teachers' decision-making and informal reasoning about culture-based socioscientific issues. Int J Sci Educ 43(5):641–671. https://doi.org/10.1080/09500693.2021.1876958

Landis J, Koch G (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174. https://doi.org/10.2307/2529310

Lind G (2006) Das Dilemma liegt im Auge des Betrachters [The dilemma is in the eye of the beholder]. PdN 55:10–16

Lind G (2019) How to teach moral competence. Logos, Berlin

Martinez M (1999) Cognition and the question of test item format. Educ Psychol 34:207–218. https://doi.org/10.1207/s15326985ep3404_2

Mathesius S, Krell M (2019) Assessing model competence with questionnaires. In: Upmeier zu Belzen A, Krüger D, van Driel J (eds) Towards a competence-based view on models and modeling in science education. Springer, Cham, pp 117–129

Messick S (1989) Validity. In: Linn RL (ed.), Educational measurement. Macmillan Publishing Co, Inc; American Council on Education, pp. 13–103

National Research Council (2001). Knowing what students know: The science and design of educational assessment. National Academies Press, Washington

Neumann K, Härtig H, Harms U, Parchmann I (2017) Science teacher preparation in Germany. In: Pedersen J, Isozaki T, Hirano T(eds) Model science teacher preparation programs. Information Age, Leeds, pp 29–52

Neumann K, Kind V, Harms U (2019) Probing the amalgam: The relationship between science teachers' content, pedagogical and pedagogical content knowledge. Int J Sci Educ 41(7):847–861. https://doi.org/10.1080/09500693.2018.1497217

Nielsen JA (2020) Teachers and socioscientific issues – an overview of recent empirical research. In: Evagorou MJA, Nielsen M, Dillon J (eds), Science teacher education for responsible citizenship. Contemporary Trends and Issues in Science Education, Vol 52. Springer, Cham

Osborne JF, Henderson JB, MacPherson A, Szu E, Wild A, Yao SY (2016) The development and validation of a learning progression for argumentation in science. J Res Sci Teach 53(6):821–846. https://doi.org/10.1002/tea.21316

Reitschert K, Langlet J, Hößle C, Mittelsten Scheid N, Schlüter K (2007) Dimensionen ethischer Urteilskompetenz: Dimensionierung und Niveaukonkretisierung [Dimensions of ethical decision-making competence]. MNU Journal 60(1):43–51

Sadler TD (2004) Informal reasoning regarding socioscientific issues. J Res Sci Teach 41: 513–536. https://doi.org/10.1002/tea.20009

Sadler TD, Donnelly LA (2006). Socioscientific Argumentation: The effects of content knowledge and morality. Int J Sci Educ, 28(12): 1463–1488. https://doi.org/10.1080/09500690600708717

Sadler TD, Fowler SR (2006) A threshold model of content knowledge transfer for socioscientific argumentation. Sci Educ 90(6):986–1004. https://doi.org/10.1002/sce.20165

Sadler TD, Zeidler DL (2005a) The significance of content knowledge for informal reasoning regarding socioscientific issues: Applying genetics knowledge to genetic engineering issues. Sci Educ, 89 (1), 71–93

Sadler TD, Zeidler DL (2005b) Patterns of informal reasoning in the context of socioscientific decision making. JRST, 42(42), 112–138

Schreier M (2012) Qualitative content analysis in practice. SAGE, London,

Sekretariat der Kultusministerkonferenz (2008) Ländergemeinsame inhaltliche Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung. [Common content requirements for subject sciences and subject didactics in teacher education], Beschluss vom 16.10.2008 i.d.F. vom 8.2.2024, Berlin

Shavelson RJ (2013) On an approach to testing and modeling competence. Educ Psychol 48:73-86. https://doi.org/10.1080/00461520.2013.779483

Steffen B, Hößle C (2014) Decision-making competence in biology education: Implementation into German curricula in relation to international approaches. J Math Sci Tech Ed 10(4):343–355. https://doi.org/10.12973/eurasia.2014.1089a

Topçu MS, Sadler TD, Yilmaz-Tüzün O (2010) Preservice science teachers' informal reasoning about socioscientific issues. Int J Sci Educ 32(18): 2475–2495. https://doi.org/10.1080/09500690903524779

Toulmin S (1958) The use of arguments. UP, Cambridge

Werner M, Schwanewedel J, Mayer J (2015) Bewertungskompetenz und der Einfluss von Kontexten und Kontext-Personen-Valenzen. [Assessment competence and the influence of contexts and context-person valences] In: Gebhard U, Hammann M, Knälmann B (eds) Bildung durch Biologieunterricht. Universität Hamburg, pp. 58–59

Yalman FE (2023) Does context affect argument quality and informal reasoning in Socio-Scientific Issues? Sci Educ Int, 34(4): 250–261. https://doi.org/10.33828/sei.v34.i4.1

Zeidler DL (2014) Socioscientific issues as a curriculum emphasis: Theory, research and practice. In: Lederman NG, Abell SK (eds.) Handbook of research on science education. Routledge NY, New York, pp. 697–726

Zuzovsky R, Donitsa-Schmidt S (2017) Comparing the effectiveness of two models of initial teacher education programmes in Israel: concurrent vs. consecutive. European Journal of Teacher Education 40(3): 413–431.