Journal of Psychoeducational Assessment
2025, Vol. 0(0) 1–17
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/07342829251364579
journals.sagepub.com/home/jpa
Sage

ASSESSRA—A Case-Based Approach for the Assessment of Students' Scientific Reasoning and Argumentation Skills

Anna Horrer[1], Laura Brandl[2], Insa Reichow[3], Michael Sailer[4], Maximilian Sailer[5], Matthias Stadler[1], Moritz Heene[2], Tamara van Gog[6], Frank Fischer[2], Martin R. Fischer[1], and Jan M. Zottmann[1]

Abstract

Scientific reasoning and argumentation (SRA) skills are crucial in higher education, yet comparing studies on these skills remains challenging due to the scarcity of well-developed SRA-tests with robust psychometric properties. In this paper, the case-based ASSESSRA approach is proposed to evaluate university students' SRA-skills, focusing specifically on the skills *evidence evaluation* and *drawing conclusions*. A prototype constructed using this approach in an educational context demonstrated reliability within an expert panel ($n = 9$; $ICC = .81$). In a subsequent study, the validity of the ASSESSRA approach was examined with 207 students, a partial-credit-model exhibited an acceptable fit, demonstrating no significant outfit and excellent distribution of ability parameters and Thurstonian thresholds. The ASSESSRA-prototype, coupled with provided guidelines, offers a versatile framework for developing comparable SRA-tests across diverse domains. This approach not only addresses the current gap in SRA assessment instruments but also holds promise for enhancing the understanding and promotion of SRA-skills in higher education.

Keywords

scientific reasoning and argumentation, case-based reasoning, assessment, evidence-based decision-making, higher education

[1]LMU University Hospital, LMU Munich, Munich, Germany
[2]LMU Munich, Munich, Germany
[3]German Research Center for Artificial Intelligence, Kaiserslautern, Germany
[4]University of Augsburg, Augsburg, Germany
[5]University of Passau, Passau, Germany
[6]Utrecht University, Utrecht, Netherlands

Corresponding Author:
Anna Horrer, Institute of Medical Education, LMU University Hospital, LMU Munich, Pettenkoferstr. 8A, 80336 Munich, Germany.
Email: anna.horrer@med.uni-muenchen.de

Scientific Reasoning and Argumentation: A Prerequisite for Evidence-Based Decision-Making

Scientific reasoning and argumentation (SRA) skills (Fischer et al., 2014) are highly valued key outcomes of (higher) education, complementing students' acquisition of subject knowledge. Across a wide range of study subjects in empirical scientific disciplines, such as psychology, medicine, or physics, SRA-skills are fundamental in the process of knowledge generation, for example, when scientists are addressing an unresolved issue, solving problems, or advancing a theory to explain unexpected results of an experiment (Heitzmann et al., 2019). SRA-skills are required of higher education graduates in both academic and professional careers, for example, when it comes to making decisions. Professional decisions are ideally based on scientifically derived evidence, that is, confirmed knowledge, rather than anecdotes, experiences, or an intuitive worldview (Harris, 2018; Rousseau, 2020; Thomas et al., 2022). Thus, SRA-skills are urgently needed in any context where the application of sound knowledge is required for evidence-based decision-making. This is the case, for instance, when medical doctors make decisions about how to treat patients, corporate managers decide on a strategy to maximize profits in their enterprise, or teachers make decisions about how to support their students to achieve particular learning outcomes. Besides professional decision-making, SRA-skills prove to be useful in order to navigate a post-truth society and to make better decisions in ambiguous situations in civic lives (Barzilai & Chinn, 2020; Scharrer et al., 2013; Zlatkin-Troitschanskaia et al., 2020). Consequently, university students are expected to develop SRA-skills during higher education to participate in the knowledge-society (Bereiter, 2002).

Notwithstanding the agreement on their relevance, SRA-skills are not clearly defined (Fischer et al., 2018). As a result, it is all the more difficult to validly assess students' SRA-skills and, ultimately, to promote them in a targeted way. Often grouped under the buzzword *21st Century Skills*, manifold conceptualizations and definitions of SRA-skills exist that are not readily distinguishable from one another (Anderman et al., 2012; Bybee, 2009). These include scientific literacy (Sadler & Zeidler, 2009), scientific thinking (Kuhn et al., 2008), argumentation in socio-scientific discourse (Chinn & Clark, 2011), critical thinking (Stanovich et al., 2013), as well as epistemic thinking or epistemic cognition (Barzilai & Weinstock, 2015; Chinn et al., 2014). The framework for scientific reasoning and argumentation by Fischer et al. (2014) assumes an interplay of eight epistemic activities in the process of SRA: *problem identification*, *questioning*, *hypothesis generation*, *construction and redesign of artefacts*, *evidence generation*, *evidence evaluation*, *drawing conclusions*, and *communicating and scrutinizing*.

Despite the diverse conceptualizations, most approaches to assessing these skills share a focus on one central research question: How do university students use evidence to make decisions in complex situations?

This question, however, is not easy to answer. To investigate how students use evidence for decision-making, it is imperative that tests to validly assess SRA-skills be developed (Daniel et al., 2019; Talman et al., 2021). Following this premise, the aim of this paper is twofold: (1) We propose the ASSESSRA approach for the development of test instruments that assess university students' SRA-skills in evaluating (scientific) evidence and in drawing evidence-based conclusions. (2) We describe the construction of an ASSESSRA-prototype in an educational context (i.e., utilizing a decision scenario from a school context) as an example. On the one hand, the psychometric properties of the prototype are shown; on the other hand, this description can serve as a blueprint for the construction of further ASSESSRA-based tests.

In light of the various shortcomings of currently available test instruments, which are discussed in greater detail below, as well as the problem of (a lack of) comparability of corresponding measures, we seek to propose an approach that has the potential to enable the construction of tests for a valid, reliable, and also comparable assessment of students' SRA-skills in different domains. In addition to construction guidelines, we also provide instructions on how to check and interpret the psychometric quality of a test constructed on the basis of the ASSESSRA approach; in this regard, we explicitly refer to the validity framework of Kane (2013). We would like to emphasize at this point that the test prototype described in this article is not a perfect instrument for assessing SRA-skills by any means—nor is it intended to be. Rather, this prototype is intended to show by way of example that the ASSESSRA approach is suitable for enabling the construction of appropriate tests.

The Need for Valid and Reliable Assessments of SRA-Skills in Higher Education

Despite the imperative for tests that validly assess SRA-skills, assessments effectively evaluating university students' SRA-skills are notably sparse (Opitz et al., 2017).

The Evidence-Based Reasoning Framework by Brown et al. (2010) offers valuable insights into evaluating students' scientific reasoning skills, particularly in science education. Notably, SRA-tests predominantly target children and students in science classrooms (Krell et al., 2022; Osborne, 2010; Osborne et al., 2013). Standardized test instruments were developed for PISA (OECD, 2019), TIMSS (Martin et al., 2016), or Lawson's Classroom Test of Scientific Reasoning (Lawson, 2004). In contrast to the abundant research on SRA-skills in primary and secondary education, higher education lacks robust assessment tools.

The assessment of SRA-skills in higher education is indispensable, influencing students' progress in the development of skills crucial for their future professions. Whether in clinical reasoning and evidence-based nursing in hospitals (Klegeris et al., 2017; Talman et al., 2021; Vierula et al., 2020) or in selecting appropriate information sources to address problematic classroom situations (Kiemer & Kollar, 2021), the need for reliable assessments is evident. However, recent reviews underscore a critical gap—studies on the development of SRA-tests in higher education are infrequent and often lack rigorous checks for psychometric properties (Kuhn et al., 2016; Opitz et al., 2017; Vierula et al., 2020).

The urgency of developing standardized SRA-tests for university students is reflected in comprehensive reviews of SRA-tests in higher education. To emphasize the urgency of developing standardized SRA-tests for university students, we distill key findings from comprehensive reviews on SRA-tests in higher education. In analyzing over 500 studies, Zlatkin-Troitschanskaia et al. (2016, 2020) provide a comprehensive overview of the research landscape on assessment instruments in higher education. While about half of the studies address interdisciplinary skills related to SRA of higher education graduates, shortcomings in existing SRA-tests are evident, as also highlighted in recent reviews of 17 SRA-tests by Talman et al. (2021) and 38 SRA-tests by Opitz et al. (2017). Criticisms from these reviews converge on the inefficiency of psychometric property checks. Moreover, the lack of objective, reliable, and valid assessment procedures for learning outcomes is prevalent. For instance, Opitz et al. (2017) criticize the absence of checks for validity and dimensionality in SRA-tests, which often rely on theoretical assumptions rather than empirical determinations, diminishing ecological validity. Talman et al. (2021) underscore that validity is infrequently verified; when reported, reliability and validity vary substantially across SRA-tests. Zlatkin-Troitschanskaia and colleagues (2016, 2020) report that many tests indirectly capture students' skills, often relying on self-reports or grades without direct reference to higher education learning outcomes.

Despite the urgency highlighted by Talman et al. (2021) in their search for high-stakes SRA-tests for student selection in higher education, the surprising finding is that there appear to be no standard tests of students' SRA-skills at present. The relevance and relationship between SRA-test scores and academic achievement remain unproven, possibly due to the scarcity of SRA-tests with high psychometric quality (Osborne, 2013; Talman et al., 2021; Zlatkin-Troitschanskaia et al., 2016). Compounded by the lack of theoretical conceptualizations of SRA-skills development, comparing the effects of studies investigating students' SRA-skills in evidence-based decision-making is a persisting challenge (Opitz et al., 2017).

This challenge is exacerbated when considering the contextual variation in SRA-tests. While Opitz et al. (2017) note that half of the SRA-tests in their review (14 out of 38) target higher education students with a predominant focus on natural sciences and medicine, Zlatkin-Troitschanskaia et al. (2016) report a concentration of studies in the context of teacher education, particularly in educational sciences or STEM subjects. The domain-specific nature of SRA-skills assessment contributes to the difficulties in comparing tests and obstructs the derivation of a unified theory on SRA-skills development.

However, two recent examples for the successful development of SRA-tests are the *Social-scientific Research Competency Test* (RCT; Gess et al., 2019) and *Leuven Research Skills Test* (LRST; Maddens et al., 2020, 2021). The RCT specifically measures qualitative and quantitative research methods knowledge of social-sciences students, whereas the LRST is targeted to assess SRA-skills of high-school students in 11th and 12th grades of a social-sciences school track. The LRST takes into account all eight epistemic activities that are theoretically assumed in the framework by Fischer et al. (2014). Both tests use case-vignettes to assess SRA-skills, the advantages of which are outlined below.

To summarize, the assessment of skills related to SRA in educational sciences has recently gained momentum (Krell et al., 2020; Zlatkin-Troitschanskaia et al., 2020). To improve professional decision-making in educational contexts, an increasing number of studies also aims at more precise SRA-tests but fail to produce reliable measures (Bicak et al., 2021). SRA-tests are typically adapted to specific study contexts, thus overarching research on the assessment of SRA-skills and the factors influencing them is inconclusive (Zlatkin-Troitschanskaia et al., 2020). In response to these challenges, we propose the ASSESSRA approach that has been explicitly designed to address the multifaceted nature of SRA-skills. ASSESSRA provides a practical means to validly assess and cultivate critical SRA-skills in university students. This paper presents construction guidelines for the ASSESSRA approach, detailing its development, steps to evaluate validity, and the potential to contribute to the assessment of university students' SRA-skills across domains.

Advantages of a Case-Based Approach for Developing SRA-Tests

With ASSESSRA, we propose a case-based approach for the development of tests to assess university students' SRA-skills *evidence evaluation* and *drawing conclusions*. This approach builds upon prior work using case-vignettes to assess SRA-skills in both educational (Engelmann et al., 2022; Trempler et al., 2015) and medical contexts (Berndt et al., 2021; Schmidt et al., 2021).

Using a case-based approach increases the ecological validity by situating the SRA-test within a credible context that mirrors real-world scenarios, such as those encountered in professional settings (Zottmann et al., 2013). Case-vignettes are regarded as suitable scaffolds in performance assessments that support test-takers to apply their knowledge and skills to a concrete scenario (Barzilai & Weinstock, 2015; Sailer et al., 2021). Secondly, case-based reasoning requires the test-takers to carry out authentic epistemic activities while processing the case information which encourages reflection and decision-making (Kolodner, 1997; Mostert, 2007). Thirdly, the use of authentic pieces of evidence increases test-takers' perceived level of appropriate difficulty of the task (Braeckman et al., 2014; Schuwirth & van der Vleuten, 2003). However, higher levels of authenticity in case-based reasoning are not linearly related to better reasoning outcomes and make the decision-making process more complex. Therefore, the inherent advantage of a case-based approach lies in its potential to bridge the gap between theoretical skill assessment and practical, real-world application, ultimately contributing to a more robust evaluation of SRA-skills.

Core Elements of an ASSESSRA-Based Test

Constructing a test based on the ASSESSRA approach involves three core elements (see Figure 1).

Decision Scenario. The case-based decision scenario featuring two potentially equivalent options to be weighed presents a dilemma wherein test-takers must opt for or against a specific solution to resolve this dilemma (Barzilai & Weinstock, 2015). After reading the scenario, the test-takers are asked to make an initial (gut) decision for or against the solution in question and to briefly justify their decision to ensure that it is rationally based and to identify any prior knowledge or biases regarding the scenario topic. To illustrate the application of the ASSESSRA approach, a prototype for the educational context was developed. The corresponding decision scenario was titled “iPads in math class” and targeted SRA-skills of educational sciences and teacher students (see Figure 2).

Standardized Abstracts. Successively, authentic pieces of evidence from meticulously selected (scientific) journal articles are presented as a sequence of standardized abstracts (see Supplemental Material for an example). The presented information for each abstract was standardized by including comparable relevant information, for example, sample-size, statistical tests such as t- or F-values, effect-size, and interpretation of the test results (if included in the original article) and length (about 300 words). If the test is computer-assisted (which is not necessarily required), the order of the standardized abstracts can be randomized. With regard to the content, the standardized abstracts deliberately present conflicting information for both sides of the dilemma instead of an unequivocal basis for decision-making in order to encourage test-takers to integrate conflicting scientific information and weigh different options before reaching a conclusion. Further, the standardized abstracts’ level of evidence strength is systematically varied. For example, article type (opinion piece, expert-interview, case study, experimental study, or meta-analysis), publication date (e.g., outdated or recent), or relevance to the decision scenario can be varied. Each standardized abstract is broken down into key statements, that is, verbatim quotations that contain

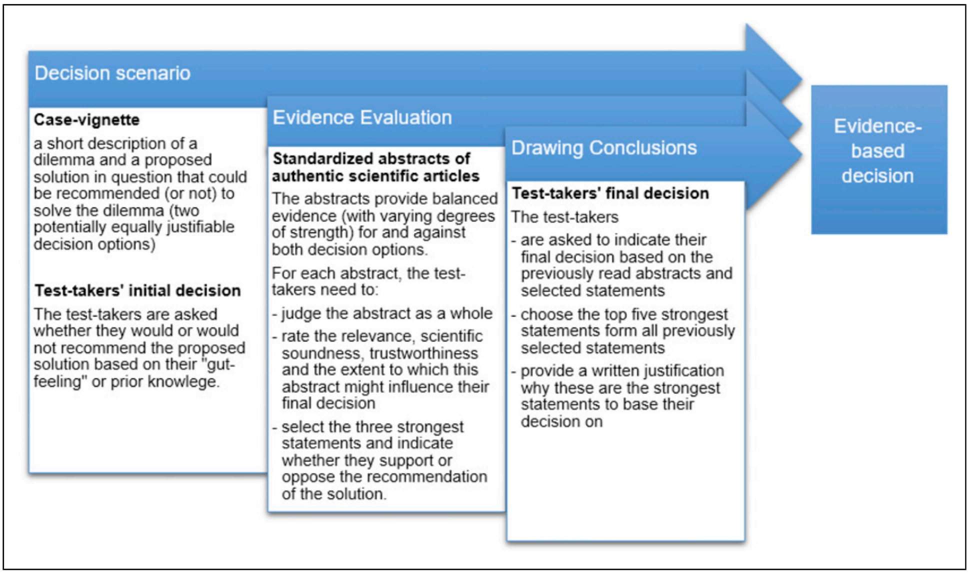
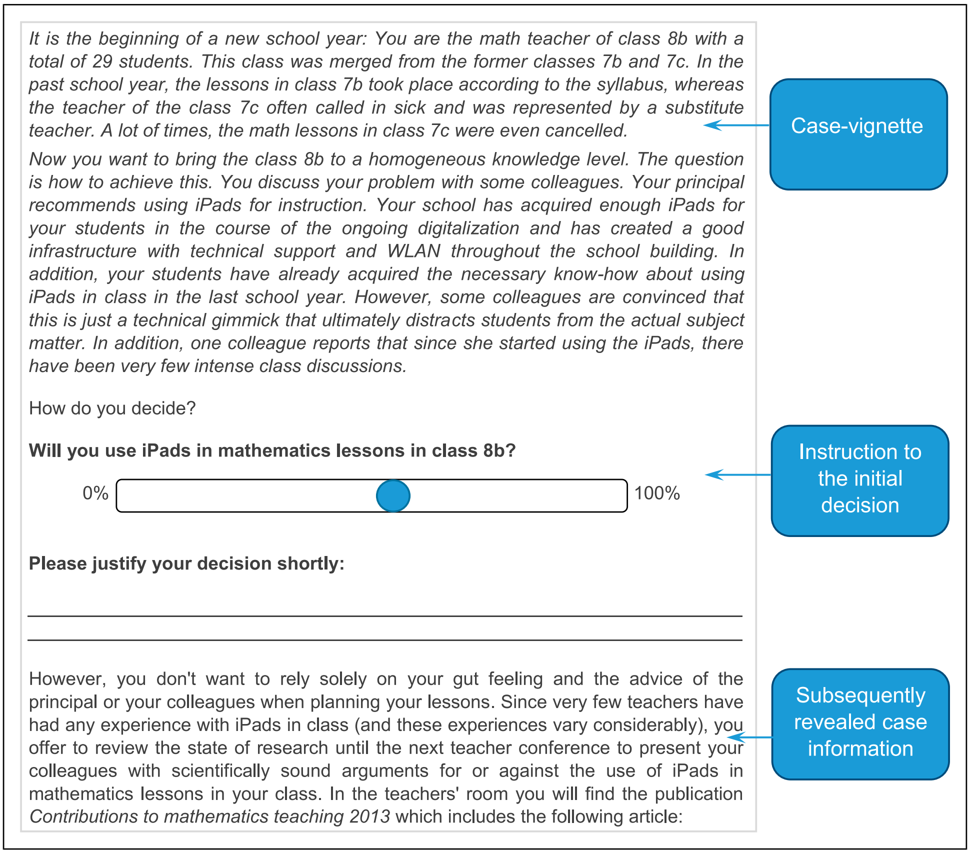

Decision scenario
Case-vignette
a short description of a dilemma and a proposed solution in question that could be recommended (or not) to solve the dilemma (two potentially equally justifiable decision options)
Test-takers' initial decision
The test-takers are asked whether they would or would not recommend the proposed solution based on their "gut-feeling" or prior knowlege.
Evidence Evaluation
Standardized abstracts of authentic scientific articles
The abstracts provide balanced evidence (with varying degrees of strength) for and against both decision options.
For each abstract, the test-takers need to:
- judge the abstract as a whole
- rate the relevance, scientific soundness, trustworthiness and the extent to which this abstract might influence their final decision
- select the three strongest statements and indicate whether they support or oppose the recommendation of the solution.
Drawing Conclusions
Test-takers' final decision
The test-takers
- are asked to indicate their final decision based on the previously read abstracts and selected statements
- choose the top five strongest statements form all previously selected statements
- provide a written justification why these are the strongest statements to base their decision on
Evidence-based decision

Figure 1. Test-takers’ process through an ASSESSRA-based test and its core elements.



It is the beginning of a new school year: You are the math teacher of class 8b with a total of 29 students. This class was merged from the former classes 7b and 7c. In the past school year, the lessons in class 7b took place according to the syllabus, whereas the teacher of the class 7c often called in sick and was represented by a substitute teacher. A lot of times, the math lessons in class 7c were even cancelled.

Now you want to bring the class 8b to a homogeneous knowledge level. The question is how to achieve this. You discuss your problem with some colleagues. Your principal recommends using iPads for instruction. Your school has acquired enough iPads for your students in the course of the ongoing digitalization and has created a good infrastructure with technical support and WLAN throughout the school building. In addition, your students have already acquired the necessary know-how about using iPads in class in the last school year. However, some colleagues are convinced that this is just a technical gimmick that ultimately distracts students from the actual subject matter. In addition, one colleague reports that since she started using the iPads, there have been very few intense class discussions.

How do you decide?

Will you use iPads in mathematics lessons in class 8b?

0% 100%

Please justify your decision shortly:

However, you don't want to rely solely on your gut feeling and the advice of the principal or your colleagues when planning your lessons. Since very few teachers have had any experience with iPads in class (and these experiences vary considerably), you offer to review the state of research until the next teacher conference to present your colleagues with scientifically sound arguments for or against the use of iPads in mathematics lessons in your class. In the teachers' room you will find the publication *Contributions to mathematics teaching 2013* which includes the following article:

Case-vignette
Instruction to the initial decision
Subsequently revealed case information

Figure 2. Case-vignette and initial decision with a short justification in the ASSESSRA-prototype.

the core information of the abstracts and represent the items in the test from which test-takers must select the strongest.

Expert-Validated Score. The score for each SRA-skill, *evidence evaluation* (EE) and *drawing conclusions* (DC), is based on the level of agreement between test-takers' selection of statements to the rating and selection of a panel of experts, comparable to the panel-based scoring method in the Script Concordance Test (Gagnon et al., 2005; Ramaekers et al., 2010). This scoring method takes into account that in the context of uncertainty, several answers might be acceptable in a reasoning process, if the expert-panel represents variability in the answers (Gagnon et al., 2011). To build meaningful scores, high agreement among the panelists is important. Gagnon et al. (2005) recommend to include at least ten experts (i.e., experienced practitioners in the domain) for acceptable reliability. The Script Concordance Test captures subjective assessments—similar to the Thurstone scale—with the aim of calculating a rate of agreement between test-takers' responses and answers given by the experts. Experts in a Script Concordance Test provide ratings for scenarios to reflect typical ways of thinking. Thurstone's scaling model allows subjective individual ranking data or comparative preference data to be converted into a single composite group interval (Krabbe, 2008).

Test Development

Empirical Example

The ASSESSRA-prototype was distributed for an expert-validation to n = 9 academic scholars (post-docs or professors) with a PhD in psychology, educational sciences, or related fields. Experts were recruited from educational sciences and educational psychology departments throughout Germany. The selected experts had strong expertise in research on scientific reasoning and argumentation. They ranged in age from 28 to 48 years, four of them were female. First, these experts evaluated each standardized abstract with a written response to determine criteria for evidence evaluation and to check whether the included abstracts covered different levels of evidence strength. Secondly, the experts formulated the strongest key messages from the respective abstracts in their own words; these were then compared with the key statements we had previously identified to ensure that all relevant statements were included. Finally, the experts rated all 60 statements (5 standardized abstracts each with 12 key statements) between 1 and 4, with 1 = simple statement without support, 2 = statement with non-empirical support (e.g., citation of a study), 3 = statement with empirical support (e.g., test of significance), and 4 = statement based on multiple empirical evidence (e.g., a meta-analysis). The score for each statement was built based on the mean value of this expert rating. Test-takers select the three strongest statements per abstract. The final score is calculated as the sum of the experts' mean ratings for these three selected statements. If always the three highest ranking statements were selected over all five abstracts, the maximum achievable score for this test prototype would be 47.

To provide evidence for the validity of the ASSESSRA-approach, we report relevant evidence to support the proposed interpretations of test scores (AERA APA NCME, 2014) following Kane's interpretation/use-argument (Kane, 2013).

Sample

Study participants were recruited via mailing lists. Educational sciences students from all semesters and all subject combinations could participate. Completion of the online questionnaire was voluntary, self-selected, and rewarded with monetary compensation. The final sample (N = 207) consisted of 102 teacher students, 57 educational sciences students, and 48 students from the master's degree program "Psychology: Learning Sciences" offered at LMU Munich. From an examination of the respective curricula and their research-related content (e.g., courses related to empirical research and work on research projects), it can be inferred that students of Psychology: Learning Sciences have the highest experience with empirical research. The mean age was 23.05 years (SD = 3.61), and 77.8 % were female.

Measures and Covariates

The online questionnaire comprised demographic and control variables, the ASSESSRA-prototype *iPads in math class*, the *Social-scientific Research Competency Test* (RCT), and *Leuven Research Skills Test* (LRST). The order of the three tests was systematically varied to compensate for sequencing effects.

EE and DC Scores. For building the EE score, test-takers must select the three strongest key statements from each standardized abstract. Each statement was given a score based on the mean of the experts' ratings. By adding the scores of the three selected statements, each standardized abstract represents a test-item. The maximum score for EE in the ASSESSRA-prototype was 47 points.

The DC-score is based on the frequency, that is, how often experts actually chose a statement to justify their final decision. After the test-takers have completed the selection of key statements, they are asked for a final decision in a scientifically sound manner by weighing the scientific evidence presented in the standardized abstracts in order to provide an *evidence-based justification* for their recommendation. For that, all previously selected statements from all standardized abstracts are presented to them again. Test-takers are then asked to indicate the five strongest statements over all 15 previously selected statements on which they would base their recommendation how to solve the dilemma. These finally selected statements build the DC-score. This score depends on how many experts chose that statement for their second decision (0 = no expert chose the statement; 0.33 = one expert chose the statement; 0.66 = two experts chose the statement; 1 = at least three experts chose the statement). The convergence of test-takers' reasoning to that of experts, that is, the evidence experts would choose to justify their decision, indicates expertise in SRA-skills.

To assess convergent validity, we compared test-takers' performance in the ASSESSRA-prototype with two established tests assessing SRA. The nine-item short version of the RCT (Gess et al., 2019) consists of the subscales knowledge of research methods, knowledge of research methodology, and research process knowledge. Each subscale addresses (1) the identification of a research problem, (2) planning a research project, and (3) analyzing and interpreting data. Each of the nine dichotomous single-choice items starts with a case-vignette and is scored either correct (1 point) or incorrect (0 points), yielding a maximum score of nine points. Wessels et al. (2020) report an acceptable reliability with a *weighted omega h* = .69.

The LRST (Maddens et al., 2020, 2021) measures all eight epistemic activities summed up to a total score by adding the average of each subscale score divided by the number of subscales. It consists of 37 items, partly in open answer or multiple-choice format. According to an extensive validation study, the LRST is suitable to assess SRA-skills and (individual differences in) overall research skills proficiency (*ordinal omega total* = .87).

Analyses

To test the degree of agreement among the expert panel's EE ratings, we computed the intraclass correlation coefficient (ICC). The value of an ICC can range from 0 to 1, with 0 indicating no reliability among raters and 1 indicating perfect reliability among raters, and values of $ICC >$ .75 indicating good reliability (Koo & Li, 2016).

To assess internal validity of the subscale EE from the ASSESSRA-prototype, we fit a partial credit item response model to the data using the TAM package (Robitzsch et al., 2022). The partial credit model (PCM), as proposed by Masters (1982), is an extension of the Rasch model, allowing items to have more than two ordered response categories. Rather than item difficulties, PCMs estimate Thurstonian thresholds that refer to the points along the latent ability trait where the probability of an individual reaching one category over another is equal (e.g., the ability at which an individual is expected to score better than the lowest score). The key indicators of interest were the reliability of the ability parameters as well as the distribution of Thurstonian thresholds and estimated EE ability scores. Model fit of the PCM was estimated using outfit, the unweighted fit mean square (Wright & Masters, 1982). Model fit was deemed acceptable if no item showed significant outfit.

Dimensionality analyses were carried out using the mirt-package (Chalmers, 2012) in R 4.4.0 (R Core Team, 2024) by comparing the fit of a two-dimensional full-information item factor-analytic generalized partial credit model (Muraki, 1992) with oblimin (i.e., oblique) rotation against a unidimensional one. To compare these models, we used Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), Sample-Size Adjusted BIC (SABIC), and Hannan–Quinn (HQ) information criterion.

To further support the validity of our test-score interpretation, we wanted to ascertain that all test-takers were already advanced in their studies and had reasonably homogeneous research skills. Thus, in the analyses with LRST, we included only students with high research experience in their study program (i.e., "Psychology: Learning Sciences" students). Educational sciences and teacher students received only the ASSESSRA-prototype and RCT. Finally, we calculated correlations between the EE and DC subscales of the ASSESSRA-prototype with LRST and RCT. Significant positive correlations would indicate convergent validity of the ASSESSRA subscales.

Results

The ICC yielded a clearly acceptable agreement value (ICC) of .812 with a 95 % confidence interval of [.722 to .878].

After removing the standardized abstract of a case study which contained mostly simple statements (scoring 1 point) and hence insufficiently captured variance in test-takers' performance, the PCM showed an acceptable fit to the data with no significant outfit for any of the items; in other words, no z-standardized outfit value exceeded |1.96|, and associated *p*-values were above 5 %, respectively (see Table 1; Tesio et al., 2024).

Reliability of the estimated EE ability scores was mediocre with a value of .62. We also assessed test information to evaluate how well the ASSESSRA-prototype differentiates test-takers at what ranges of ability. Both ability parameters and Thurstonian thresholds showed a very good distribution with showing neither substantial shift nor critical outliers (Figure 3).

Results of dimensionality analyses are displayed in Table 2. As indicated by the consistently smaller penalty function values, the unidimensional model fitted relatively better than the two-dimensional model. Table 3 shows the standardized factor loadings of the five EE items on the factor. Notice that IRT-related item slope parameters were transformed into factor loadings for ease of interpretation. For clarity of interpretation, we converted item slopes $\boldsymbol{\alpha}_{\mathrm{k}}$ of items $i = 1,\ldots,k$ into more easily interpretable standardized factor loadings $\boldsymbol{\lambda}_{\mathrm{k}}$ using the transformation formula $\lambda_i = \frac{\alpha_i}{\sqrt{1+\alpha_i^2}}$ (e.g., Gibbons et al., 2014).

All items showed positive associations with the latent factor, although it has to be stated that the loadings were relatively small, given that factor loadings above 0.36 are considered moderate. However, in larger samples, even small factor loadings can be considered significant (Stevens, 2001). While small factor loadings are not uncommon for affective or socio-cognitive constructs (e.g., Jordan & Spiess, 2019), the relatively low factor loadings indicate room for improvement regarding item quality.

Table 1. Outfit Statistics for the PCM.

Item	Outfit	*Zstd Outfit*	*p*
EE_Text2	0.98	−0.15	.883
EE_Text3	0.92	−0.77	.444
EE_Text4	1.02	0.20	.842
EE_Text5	1.04	0.43	.665

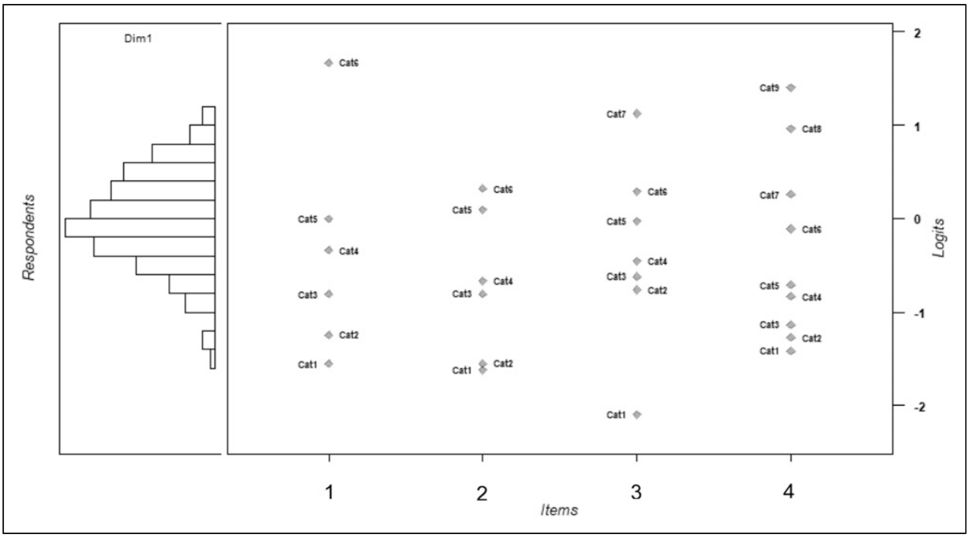


Figure 3. Wright map visualizing ability estimates and Thurstonian thresholds for the PCM.

Table 2. IRT Model Fit Comparison.

	AIC	SABIC	HQ	BIC
Unidimensional	3096.75	3102.50	3143.92	3213.39
Two-dimensional	3103.10	3109.51	3155.67	3233.08

Table 3. Factor Loading of the EE Items.

	Factor Loading
EE_Text1	0.26
EE_Text2	0.34
EE_Text3	0.44
EE_Text4	0.26
EE_Text5	0.20

Finally, we calculated correlations between the EE and DC subscales of the ASSESSRA-prototype with the LRST and the RCT to generate evidence on the measure's convergent validity. For the EE subscale, we found a moderate correlation (according to the classification of Cohen, 1988) with LRST ($n = 48$, $r = .30$, $p = .021$) and a small correlation with RCT ($n = 207$, $r = .18$, $p = .006$). For the DC subscale, we found a moderate correlation with LRST ($n = 48$, r = .28, $p = .028$) and a small correlation with RCT ($n = 207$, $r = .16$, $p = .010$).

Discussion

In this paper, we have introduced the case-based ASSESSRA approach for the construction of SRA-tests for university students and exemplarily examined psychometric properties for an ASSESSRA-prototype.

The high convergence of experts' ratings supports the reliability assumption of the ASSESSRA-prototype. Although the low reliability in *evidence evaluation* of the student sample seems to contrast with the substantially high intra-class correlation of the experts, the good distribution of both ability parameters and Thurstonian thresholds indicates that the ASSESSRA-prototype was suitable to capture SRA-skills in students with different abilities. Taking the performance of experts and students into account, constructing a reliable SRA-test for university students can be considered achieved. The construct validity of the ASSESSRA-prototype, which aims to assess university students' scientific reasoning and argumentation (SRA) skills, received mixed support from our analysis. The correlations with two previously validated tests, RCT and LRST, were mediocre to rather low. Notably, the subscales *evidence evaluation* and *drawing conclusions* showed higher correlations with the LRST, which is grounded in the same theoretical framework by Fischer et al. (2014). These moderate correlations suggest convergent validity for assessing SRA, considering that the LRST aggregates an overall SRA value across eight epistemic activities, whereas the ASSESSRA-based test focuses on two specific activities. The lower correlations between ASSESSRA subscales and the RCT were anticipated. The RCT explicitly assesses knowledge of methods, methodologies, and research processes, while ASSESSRA implicitly evaluates this knowledge through students' performance in reasoning with methodologically sound evidence. These nuanced differences likely contributed to the suppressed correlations.

Applying Kane's framework of validity arguments (Kane, 2013), we evaluated the ASSESSRA-prototype across the dimensions of scoring, generalization, extrapolation, and implications (Cook et al., 2015). The scoring dimension examines the construction of specific items to build scores. Our findings revealed ceiling effects, similar to issues identified in other measurement instruments (Bao et al., 2022), suggesting an imprecise measurement of evidence evaluation. Additionally, the sum score calculation for drawing conclusions depends heavily on the response behavior of the expert panel. Future research should address optimal panel composition to ensure reproducible and reliable values (Gagnon et al., 2005). Generalization considers how well test items represent the broader test universe. The ASSESSRA approach uses standardized abstracts from authentic sources, adaptable to diverse contexts, supporting the argument that the test items can be generalized. The core elements, such as case vignettes, standardized abstracts, and expert-validated scores, can be modified and refined as necessary, reinforcing the generalizability of the ASSESSRA approach. Extrapolation involves using test scores as a representation of real-world performance. ASSESSRA-based tests, designed with authentic evidence and a case-based approach, ensure high content validity, indicating that test performance reflects real-world SRA. This supports the extrapolation validity of the ASSESSRA approach. Implications refer to the application of test scores for decisions, such as pass/fail outcomes. While our study suggests reasonable evidence of validity in scoring, generalization, and extrapolation, further research is needed across different domains and scenarios to evaluate the ASSESSRA approach for high-stakes purposes, such as student admissions (Talman et al., 2021).

In summary, while the ASSESSRA-prototype shows promise in measuring SRA, particularly in terms of generalization and extrapolation, certain aspects, particularly scoring, require further investigation and refinement. Future studies should focus on addressing these issues to enhance the overall validity and reliability of the ASSESSRA approach.

Limitations

Although we argue that ASSESSRA in general is an innovative approach to develop tests for the valid assessment of students' SRA-skills, it is difficult to ensure robust psychometric properties of tests used for students—as also pointed out by previous reviews on SRA-tests (Opitz et al., 2017; Talman et al., 2021; Zlatkin-Troitschanskaia et al., 2016). The result of mean correlation values does not represent strong evidence for or against the assumption of construct validity. Potentially, a different, unrelated variable could have been selected that would have led to correlations of a comparable magnitude. One possible reason for this could have been the mediocre reliability of the ASSESSRA-prototype, which could limit the maximum observable correlation (however, cf. Edelsbrunner, 2022; Stadler et al., 2021). We emphasize that every test built according to the ASSESSRA approach must be thoroughly checked for psychometric quality.

Conclusion and Outlook

In the evolving landscape of higher education, the assessment of SRA skills remains a major challenge, since traditional assessments such as multiple-choice questions or written exams may not fully encompass the dynamic and process-oriented nature of SRA (Barzilai & Weinstock, 2015). That is because the development of instruments to assess cognitive skills requires complex performance-based methods, which is more challenging than the assessment of subject knowledge (C. Kuhn et al., 2016; Shavelson, 2013). Moreover, existing assessment instruments often fall short of capturing the nuanced dimensions of SRA or have limitations in terms of meeting quality criteria (Bao et al., 2022; Opitz et al., 2017; Talman et al., 2021).

Our aim with ASSESSRA was to provide useful guidelines for the construction of SRA-tests. The test prototype described here (and the steps described for gathering evidence *to support the argument of validity and reliability*) should be seen as a blueprint for the construction of further ASSESSRA-based tests. In our view, the ASSESSRA approach can contribute to further developing the quality of current assessments of SRA skills within a common framework. The next step will be to investigate whether our approach can be used to develop tests for more domains in which students' SRA-skills are tested; for instance, it is currently being used to develop an instrument to assess SRA-skills in medical education. Further studies need to examine whether an ASSESSRA-based test might be sensitive enough to distinguish SRA-skills of different student groups *within* a given domain. Additionally, the question remains open if and how results of ASSESSRA-based tests could be made comparable *across* domains. It may be possible to develop specific scenarios for this purpose that are equally suitable for testing students from different domains (Berndt et al., 2021).

To derive a theory about SRA-skills development in higher education, there is a strong need to ensure better comparability of test instruments and to identify and subsequently control for factors that may influence SRA (e.g., epistemological beliefs, cognitive rigidity, cognitive load, or intelligence). In the long term, this might add to the empirical derivation of a theory of university students' SRA-skills development and its influencing factors across a broad range of different domains. In addition to an evaluative assessment of individual student SRA-skills, ASSESSRA-based tests could also be used in university teaching to stimulate reflection on the value of research findings and methods in one's study domain, thus promoting engagement in SRA.

Acknowledgments

The authors would like to thank Louise Maddens and our expert panel for their ratings, comments, and suggestions to the present study. We thank our student researchers Tolgonai Erkinova, Annabell Reinhardt, and Ruyi Zhao for their help during data collection.

ORCID iDs

Anna Horrer https://orcid.org/0000-0003-1029-5896
Laura Brandl https://orcid.org/0000-0001-7974-7892
Insa Reichow https://orcid.org/0000-0002-7799-8409
Maximilian Sailer https://orcid.org/0000-0003-2241-7694
Matthias Stadler https://orcid.org/0000-0001-8241-8723
Moritz Heene https://orcid.org/0009-0007-4956-2417
Tamara van Gog https://orcid.org/0000-0003-3766-6255
Frank Fischer https://orcid.org/0000-0003-0253-659X
Martin R. Fischer https://orcid.org/0000-0002-5299-5025
Jan M. Zottmann https://orcid.org/0000-0002-3887-1181

Ethical Approval

The LMU Ethics Committee approved the study (project number 20-699).

Informed Consent from Participants

Informed consent was obtained from all individual participants included in the study.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Federal Ministry of Education and Research (grant no. 01PB18004C) and the Elite Network Bavaria (grant no. K-GS-2012-209).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest concerning the research, authorship, and/or publication of this article.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Supplemental Material

Supplemental material for this article is available online.

References

AERA APA NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association. https://www.apa.org/science/programs/testing/standards

Anderman, E. M., Sinatra, G. M., & Gray, D. L. (2012). The challenges of teaching and learning about science in the twenty-first century: Exploring the abilities and constraints of adolescent learners. *Studies in Science Education*, *48*(1), 89–117. https://doi.org/10.1080/03057267.2012.655038

Bao, L., Koenig, K., Xiao, Y., Fritchman, J., Zhou, S., & Chen, C. (2022). Theoretical model and quantitative assessment of scientific thinking and reasoning. *Physical Review Physics Education Research*, *18*(1), 010115. https://doi.org/10.1103/PhysRevPhysEducRes.18.010115

Barzilai, S., & Chinn, C. A. (2020). A review of educational responses to the "post-truth" condition: Four lenses on "post-truth" problems. *Educational Psychologist*, *55*(3), 107–119. https://doi.org/10.1080/00461520.2020.1786388

Barzilai, S., & Weinstock, M. (2015). Measuring epistemic thinking within and across topics: A scenario-based approach. *Contemporary Educational Psychology*, *42*(1), 141–158. https://doi.org/10.1016/j.cedpsych.2015.06.006

Bereiter, C. (2002). *Education and mind in the knowledge age*. L. Erlbaum Associates. https://doi.org/10.4324/9781410612182

Berndt, M., Schmidt, F., Sailer, M., Fischer, F., Fischer, M. R., & Zottmann, J. (2021). Investigating statistical literacy and scientific reasoning & argumentation in medical-social sciences-and economics students. *Learning and Individual Differences*, *86*(2), Article 101963. https://doi.org/10.1016/j.lindif.2020.101963

Bicak, B. E., Borchert, C. E., & Höner, K. (2021). Measuring and fostering preservice chemistry teachers' scientific reasoning competency. *Education Sciences*, *11*(9), 496. https://doi.org/10.3390/educsci11090496

Braeckman, L., 't Kint, L., Bekaert, M., Cobbaut, L., & Janssens, H. (2014). Comparison of two case-based learning conditions with real patients in teaching occupational medicine. *Medical Teacher*, *36*(4), 340–346. https://doi.org/10.3109/0142159X.2014.887833

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

Bybee, R. W. (2009). The BSCS 5E instructional model and 21st century skills: A commissioned paper prepared for a workshop on exploring the intersection of science education and the development of 21st century. https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_073327.pdf

Chalmers, R. P. (2012). Mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chinn, C. A., & Clark, D. B. (2011). Learning through collaborative argumentation. In *The international handbook of collaborative learning* (pp. 314–332). Taylor & Francis. https://doi.org/10.4324/9780203837290.ch18

Chinn, C. A., Rinehart, R., & Buckland, L. (2014). Epistemic cognition and evaluating information: Applying the AIR model of epistemic cognition. In D. Rapp & J. L. G. Braasch (Eds.), *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences* (pp. 425–453). The MIT Press. https://doi.org/10.7551/mitpress/9737.003.0025

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=1192162

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education*, *49*(6), 560–575. https://doi.org/10.1111/medu.12678

Daniel, M., Rencic, J., Durning, S. J., Holmboe, E., Santen, S. A., Lang, V., Ratcliffe, T., Gordon, D., Heist, B., Lubarsky, S., Estrada, C. A., Ballard, T., Artino, A. R., Da Sergio Silva, A., Cleary, T., Stojan, J., & Gruppen, L. D. (2019). Clinical reasoning assessment methods: A scoping review and practical guidance. *Academic Medicine: Journal of the Association of American Medical Colleges*, *94*(6), 902–912. https://doi.org/10.1097/ACM.0000000000002618

Edelsbrunner, P. A. (2022). A model and its fit lie in the eye of the beholder: Long live the sum score. *Frontiers in Psychology*, *13*, Article 986767. https://doi.org/10.3389/fpsyg.2022.986767

Engelmann, K., Hetmanek, A., Neuhaus, B. J., & Fischer, F. (2022). Testing an intervention of different learning activities to support students' critical appraisal of scientific literature. *Frontiers in Education*, *7*, Article 977788. https://doi.org/10.3389/feduc.2022.977788

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

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., Neuhaus, B., Dorner, B., Pankofer, S., Fischer, M., Strijbos, J.-W., Heene, M., & Eberle, J. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research*, *2*(3), 28–45. https://doi.org/10.14786/flr.v2i3.96

Gagnon, R., Charlin, B., Coletti, M., Sauvé, E., & van der Vleuten, C. (2005). Assessment in the context of uncertainty: How many members are needed on the panel of reference of a script concordance test? *Medical Education*, *39*(3), 284–291. https://doi.org/10.1111/j.1365-2929.2005.02092.x

Gagnon, R., Lubarsky, S., Lambert, C., & Charlin, B. (2011). Optimization of answer keys for script concordance testing: Should we exclude deviant panelists, deviant responses, or neither? *Advances in Health Sciences Education: Theory and Practice*, *16*(5), 601–608. https://doi.org/10.1007/s10459-011-9279-2

Gess, C., Geiger, C., & Ziegler, M. (2019). Social-scientific research competency: Validation of test score interpretations for evaluative purposes in higher education. *European Journal of Psychological Assessment*, *35*(5), 737–750. https://doi.org/10.1027/1015-5759/a000451

Gibbons, R. D., Perraillon, M. C., & Kim, J. B. (2014). Item response theory approaches to harmonization and research synthesis. *Health Services & Outcomes Research Methodology*, *14*(4), 213–231. https://doi.org/10.1007/s10742-014-0125-x

Harris, K. R. (2018). Educational psychology: A future retrospective. *Journal of Educational Psychology*, *110*(2), 163–173. https://doi.org/10.1037/edu0000267

Heitzmann, N., Seidel, T., Hetmanek, A., Wecker, C., Fischer, M. R., Ufer, S., Schmidmaier, R., Neuhaus, B., Siebeck, M., Stürmer, K., Obersteiner, A., Reiss, K., Girwidz, R., Fischer, F., & Opitz, A. (2019). Facilitating diagnostic competences in simulations in higher education A framework and a research agenda. *Frontline Learning Research*, *7*(4), 1–24. https://doi.org/10.14786/flr.v7i4.384

Jordan, P., & Spiess, M. (2019). Rethinking the interpretation of item discrimination and factor loadings. *Educational and Psychological Measurement*, *79*(6), 1103–1132. https://doi.org/10.1177/0013164419843164

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kiemer, K., & Kollar, I. (2021). Source selection and source use as a basis for evidence-informed teaching. *Zeitschrift für Pädagogische Psychologie*, *35*(2–3), 127–141. https://doi.org/10.1024/1010-0652/a000302

Klegeris, A., McKeown, S. B., Hurren, H., Spielman, L. J., Stuart, M., & Bahniwal, M. (2017). Dynamics of undergraduate student generic problem-solving skills captured by a campus-wide study. *Higher Education*, *74*(5), 877–896. https://doi.org/10.1007/s10734-016-0082-0

Kolodner, J. L. (1997). A view from case-based reasoning. *The American Psychologist*, *52*(1), 57–66. https://doi.org/10.1037//0003-066x.52.1.57

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Krabbe, P. F. M. (2008). Thurstone scaling as a measurement method to quantify subjective health outcomes. *Medical Care*, *46*(4), 357–365. https://doi.org/10.1097/MLR.0b013e31815ceca9

Krell, M., Redman, C., Mathesius, S., Krüger, D., & van Driel, J. (2020). Assessing pre-service science teachers' scientific reasoning competencies. *Research in Science Education*, *50*(6), 2305–2329. https://doi.org/10.1007/s11165-018-9780-1

Krell, M., Vorholzer, A., & Nehring, A. (2022). Scientific reasoning in science education: From global measures to fine-grained descriptions of students' competencies. *Education Sciences*, *12*(2), 97. https://doi.org/10.3390/educsci12020097

Kuhn, C., Zlatkin-Troitschanskaia, O., Pant, H. A., & Hannover, B. (2016). Valide Erfassung der Kompetenzen von Studierenden in der Hochschulbildung. *Zeitschrift für Erziehungswissenschaft*, *19*(2), 275–298. https://doi.org/10.1007/s11618-016-0673-7

Kuhn, D., Iordanou, K., Pease, M., & Wirkala, C. (2008). Beyond control of variables: What needs to develop to achieve skilled scientific thinking? *Cognitive Development*, *23*(4), 435–451. https://doi.org/10.1016/j.cogdev.2008.09.006

Lawson, A. E. (2004). The nature and development of scientific reasoning: A synthetic view. *International Journal of Science and Mathematics Education*, *2*(3), 307–338. https://doi.org/10.1007/s10763-004-3224-2

Maddens, L., Depaepe, F., Janssen, R., Raes, A., & Elen, J. (2020). Evaluating the leuven research skills test for 11th and 12th grade. *Journal of Psychoeducational Assessment*, *38*(4), 445–459. https://doi.org/10.1177/0734282918825040

Maddens, L., Depaepe, F., Janssen, R., Raes, A., & Elen, J. (2021). Research skills in upper secondary education and in first year of university. *Educational Studies*, *47*(4), 491–507. https://doi.org/10.1080/03055698.2020.1715204

Martin, M. O., Mullis, I. V. S., Foy, P., & Hooper, M. (2016). TIMSS 2015 international results in science. https://timss2015.org/wp-content/uploads/filebase/fullpdfs/T15-International-Results-in-Science-Grade-4.pdf

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/bf02296272

Mostert, M. P. (2007). Challenges of case-based teaching. *The Behavior Analyst Today*, *8*(4), 434–442. https://doi.org/10.1037/h0100632

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. https://doi.org/10.1177/014662169201600206

OECD. (2019). PISA 2018 Ergebnisse (Band I). https://doi.org/10.1787/1da50379-de

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

Osborne, J. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *Science*, *328*(5977), 463–466. https://doi.org/10.1126/science.1182595

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

Osborne, J., Simon, S., Christodoulou, A., Howell-Richardson, C., & Richardson, K. (2013). Learning to argue: A study of four schools and their attempt to develop the use of argumentation as a common instructional practice and its impact on students. *Journal of Research in Science Teaching*, *50*(3), 315–347. https://doi.org/10.1002/tea.21073

Ramaekers, S., Kremer, W., Pilot, A., van Beukelen, P., & van Keulen, H. (2010). Assessment of competence in clinical reasoning and decision-making under uncertainty: The script concordance test method. *Assessment & Evaluation in Higher Education*, *35*(6), 661–673. https://doi.org/10.1080/02602938.2010.500103

R Core Team. (2024). R: A language and environment for statistical computing. https://www.R-project.org/

Robitzsch, A., Kiefer, T., & Wu, M. (2022). TAM: Test analysis modules. R package version. https://CRAN.R-project.org/package=TAM

Rousseau, D. M. (2020). Making evidence-based organizational decisions in an uncertain world. *Organizational Dynamics*, *49*(1), Article 100756. https://doi.org/10.1016/j.orgdyn.2020.100756

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

Sailer, M., Stadler, M., Schultz-Pernice, F., Franke, U., Schöffmann, C., Paniotova, V., Husagic, L., & Fischer, F. (2021). Technology-related teaching skills and attitudes: Validation of a scenario-based selfassessment instrument for teachers. *Computers in Human Behavior*, *115*(2), Article 106625. https://doi.org/10.1016/j.chb.2020.106625

Scharrer, L., Britt, M. A., Stadtler, M., & Bromme, R. (2013). Easy to understand but difficult to decide: Information comprehensibility and controversiality affect laypeople's science-based decisions. *Discourse Processes*, *50*(6), 361–387. https://doi.org/10.1080/0163853X.2013.813835

Schmidt, F., Zottmann, J., Sailer, M., Fischer, M. R., & Berndt, M. (2021). Statistical literacy and scientific reasoning & argumentation in physicians. *GMS Journal for Medical Education*, *38*(4), Doc77. https://doi.org/10.3205/ZMA001473

Schuwirth, L., & van der Vleuten, C. (2003). The use of clinical simulations in assessment. *Medical Education*, *37*(Suppl. 1), 65–71. https://doi.org/10.1046/j.1365-2923.37.s1.8.x

Shavelson, R. J. (2013). On an approach to testing and modeling competence. *Educational Psychologist*, *48*(2), 73–86. https://doi.org/10.1080/00461520.2013.779483

Stadler, M., Sailer, M., & Fischer, F. (2021). Knowledge as a formative construct: A good alpha is not always better. *New Ideas in Psychology*, *60*(1), Article 100832. https://doi.org/10.1016/j.newideapsych.2020.100832

Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science*, *22*(4), 259–264. https://doi.org/10.1177/0963721413480174

Stevens, J. P. (2001). *Applied multivariate statistics for the social sciences*. Psychology Press. https://doi.org/10.4324/9781410604491

Talman, K., Vierula, J., Kanerva, A.-M., Virkki, O., Koivisto, J.-M., & Haavisto, E. (2021). Instruments for assessing reasoning skills in higher education: A scoping review. *Assessment & Evaluation in Higher Education*, *46*(3), 376–392. https://doi.org/10.1080/02602938.2020.1776212

Tesio, L., Caronni, A., Simone, A., Kumbhare, D., & Scarano, S. (2024). Interpreting results from Rasch analysis 2. Advanced model applications and the data-model fit assessment. *Disability and Rehabilitation*, *46*(3), 604–617. https://doi.org/10.1080/09638288.2023.2169772

Thomas, A., Chin-Yee, B., & Mercuri, M. (2022). Thirty years of teaching evidence-based medicine: Have we been getting it all wrong? *Advances in Health Sciences Education*, *27*, 263–376. Advance online publication. https://doi.org/10.1007/s10459-021-10077-4

Trempler, K., Hetmanek, A., Wecker, C., Kiesewetter, J., Wermelt, M., Fischer, F., Fischer, M., & Gräsel, C. (2015). *Nutzung von Evidenz im Bildungsbereich. Validierung eines Instruments zur Erfassung von Kompetenzen der Informationsauswahl und Bewertung von Studien*. Beltz Juventa. https://doi.org/10.25656/01:15508

Vierula, J., Haavisto, E., Hupli, M., & Talman, K. (2020). The assessment of learning skills in nursing student selection: A scoping review. *Assessment & Evaluation in Higher Education*, *45*(4), 496–512. https://doi.org/10.1080/02602938.2019.1666970

Wessels, I., Rueß, J., Gess, C., Deicke, W., & Ziegler, M. (2020). Is research-based learning effective? Evidence from a pre–post analysis in the social sciences. *Studies in Higher Education*, *46*(1), 1–15. https://doi.org/10.1080/03075079.2020.1739014

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA Press.

Zlatkin-Troitschanskaia, O., Pant, H. A., Toepper, M., & Lautenbach, C. (2020). *Student learning in German higher education*. Springer. https://doi.org/10.1007/978-3-658-27886-1

Zlatkin-Troitschanskaia, O., Pant, H. A., Kuhn, C., Toepper, M., & Lautenbach, C. (2016). *Messung akademisch vermittelter Kompetenzen von Studierenden und Hochschulabsolventen: Ein Überblick zum nationalen und internationalen Forschungsstand. Edition ZfE: Band 1*. Springer VS. https://doi.org/10.1007/978-3-658-10830-4

Zottmann, J., Stegmann, K., Strijbos, J.-W., Vogel, F., Wecker, C., & Fischer, F. (2013). Computer-supported collaborative learning with digital video cases in teacher education: The impact of teaching experience on knowledge convergence. *Computers in Human Behavior*, *29*(5), 2100–2108. https://doi.org/10.1016/j.chb.2013.04.014