

Article

Climbing the STAIRs: Assessing students' social scientific reasoning skills

Thomas Klijnstra¹, Geerte Savenije¹, Chiel Huijskes² & Carla van Boxtel¹

¹University of Amsterdam, Amsterdam, The Netherlands

²National Institute for Educational Measurement, Arnhem, The Netherlands

Highlights:

- Assessing complex skills in secondary school teaching practice is considered challenging.
- We developed items (STAIRs) to formatively assess students' social scientific reasoning.
- STAIRs were validated by experts, teachers, think-aloud interviews, and test administration.
- STAIRs elicited students' reasoning about social problems in three proficiency levels.
- The design principles may be applied by teachers in the development of assessment items.

Purpose: Assessing complex skills is considered important but challenging. This study focused on developing assessment items to evaluate secondary social science students' proficiency in the sub-skill of causal analysis.

Design/methodology/approach: Based on a conceptual framework of social scientific reasoning, we designed formative assessment items known as STAIRs (Social science Teaching Assessment Items of Reasoning). The STAIRs were validated in three focus groups: two groups of assessment experts ($N = 7$ and $N = 3$) and one group of social science teachers ($N = 10$). Additionally, think-aloud interviews were conducted with eight social science students. The quality of the STAIRs was evaluated by administering the items to 338 social science students in 21 Dutch social science classes.

Findings: The results showed that it is possible to distinguish between the three performance levels in students' reasoning using the STAIRs.

Practical implications: The design principles for the STAIRs may aid teachers in developing additional assessment items.

Keywords: social science education, formative assessment, assessment design, social scientific reasoning, causal analysis, civic education

Corresponding author:

Thomas Klijnstra, Research Institute of Child Development and Education (RICDE), PO Box 15776, 1001 NG, Amsterdam, The Netherlands.

E-Mail: t.klijnstra@uva.nl

Suggested citation:

Klijnstra, Thomas, Savenije, Geerte, Huijskes, Chiel, & van Boxtel, Carla (2025). Climbing the STAIRs: Assessing students' social scientific reasoning skills. *Journal of Social Science Education*, 24(2).
<https://doi.org/10.11576/jsse-7938>

 Open access



1 INTRODUCTION

Reasoning about social problems is considered an essential and relevant skill, both for the individual and for a democratic society (Abrami et al., 2008; Lee et al., 2021). A key objective of social science education is to enable students to critically analyse and reason about social problems, such as climate change, youth crime, and social inequality (Klijnstra et al., 2023; Sandahl, 2015). In the Dutch social science program for upper secondary education, the concept–context approach is central: Students must learn to use social scientific concepts to analyse social problems that function as context (College van Toetsen en Examens, 2019; Klijnstra et al., 2023; Olgers et al., 2021).

Although the importance of teaching complex skills is widely recognised in the educational field (Brookhart, 2010; Ercikan & Seixas, 2015; Schraw & Robinson, 2011), assessing these complex skills remains challenging (Ercikan & Seixas, 2015). Complex skills are often assessed through large, authentic take-home assignments (Ercikan & Seixas, 2015; Lee et al., 2021). However, teachers also require smaller tasks to evaluate students' reasoning about social problems in a more formative manner. This need is also evident in the assessment of complex skills within social science education and civics (Jansson, 2023).

In this study, we have designed assessment items aimed at formatively evaluating students' social scientific reasoning abilities, which we have defined as STAIRs (Social science Teaching Assessment Items of Reasoning). The research question guiding this study is: *What types of short-answer questions can serve as valid (formative) tools for assessing students' social scientific reasoning?* Ultimately, social science teachers can use these STAIRs to gain more insights into students' proficiency levels in social scientific reasoning.

2 THEORETICAL FRAMEWORK

2.1 Domain analysis: Operationalisation of social scientific reasoning

The operationalisation of specific skills is a logical first step in principled design approaches that focus on constructing assessment items for complex skills (Breakstone, 2014; Löfström et al., 2023; Messick, 1994; Mislevy et al., 2003; Pellegrino et al., 2001; Schmeiser & Welch, 2006). Therefore, in line with two crucial principled design approaches – the Assessment Triangle (Pellegrino et al., 2001) and the Evidence-Centered Design model (Mislevy et al., 2003) – it is essential to define and operationalise the specific domain of social scientific reasoning. More specifically, it is crucial to operationalise the specific knowledge, skills, or attitudes that need to be assessed (Messick, 1994; Mislevy et al., 2003; Pellegrino et al., 2001), thereby providing teachers with greater clarity regarding what they and their students need to know or do (Löfström & Ouakrim-Soivio, 2022; Wiliam & Leahy, 2015).

If academic literature on social science education and curriculum documents of policy makers lack an operationalisation of social science reasoning in curriculum documents, it is unsurprising that social science teachers experience difficulties in the teaching and assessing of students' social scientific reasoning (Jansson, 2023; Klijnstra et al., 2023; Van Boxtel et al., 2017). In a previous study, we conceptualised and operationalised students' social scientific reasoning (Klijnstra et al., 2023). Building on conceptualisations in academic handbooks of sociology and political science (e.g., Ultee et al., 2003; Van Tubergen, 2020; Woerdman, 2013), social science education (e.g., Sandahl, 2015), civic reasoning (e.g., Lee et al., 2021), and historical reasoning (e.g., Seixas et al., 2013; Van Boxtel & Van

Drie, 2018), and based on an analysis of student papers in social science classes, we operationalised the domain of students' social scientific reasoning. This operationalisation identified five main reasoning activities: (1) causal analysis; (2) use of social scientific concepts, models, and theories; (3) use of evidence; (4) use of perspectives and reflection on them; and (5) comparing.

Furthermore, we operationalised each reasoning activity into subcategories (e.g., distinguishing causes and consequences) and divided each subcategory into three proficiency levels of social scientific reasoning, supported by practical examples of students' reasoning. This operationalisation of social scientific reasoning can serve as a crucial first step in developing assessment items, such as STAIRs, which evaluate subskills of students' social scientific reasoning.

2.2 Exploring ways of assessing social scientific reasoning

As previously mentioned, teachers find it challenging to assess complex skills (Brookhart, 2010; Ercikan & Seixas, 2015). This challenge extends to social science and civics teachers: although the teaching of reasoning has gained more attention in teaching practices, teachers still struggle to develop assessments that effectively elicit complex skills in these subjects (Jansson, 2023; Lee et al., 2021; Sluijsmans, 2013, 2014; Van Boxtel et al., 2017). Most standardised assessments within civics teaching primarily focus on the recall of factual knowledge rather than performance-based assessments (Brookhart & Durkin, 2003; Curry & Smith, 2017). Therefore, in line with insights from research on other subjects, teachers require advanced knowledge regarding the design and evaluation of assessments that elicit complex skills, and further professional development is necessary (Amani et al., 2021; Campbell, 2013; Cooper et al., 2017; Jansson, 2023; Lee et al., 2021; Moss, 2013; Sluijsmans, 2013, 2014).

When complex skills are assessed in social science education and civics, they are often evaluated through more extensive written assignments (Jansson, 2023) and authentic learning tasks (Van Boxtel et al., 2017). Generally, authentic assessment tasks help students make abstract social science concepts more meaningful and explore the application of knowledge and skills to real-world problems (Bransford et al., 2000; Breakstone, 2014; Brown et al., 1989; Jansson, 2023; Maddox & Saye, 2017; Newmann et al., 2016). These tasks are typically extensive and require significant time to complete, and they can be administered as individual or group assignments (Maddox & Saye, 2017). Examples of authentic tasks in social science education include policy papers and advice letters to municipal councils addressing the reception of refugees, as well as inquiry tasks focused on social inequality and youth obesity.

An advantage of open-ended authentic learning tasks is that they facilitate the transfer of more abstract concepts to students' real-world problems and can motivate and facilitate their complex reasoning skills (Maddox & Saye, 2017). However, these tasks are also time-consuming for teachers to develop and for students to complete (Ercikan & Seixas, 2015). Furthermore, due to the complexity of the tasks, such as measuring multiple skills within a single assignment, it becomes challenging for both teachers and students to evaluate students' proficiency in a single or limited number of skills (Smith & Breakstone, 2015; Young & Leinhardt, 1998). This is particularly important for formative assessment, where identifying students' ability to demonstrate certain components of social scientific reasoning and providing targeted feedback is essential.

In contrast, multiple-choice assessments are less time-consuming for teachers to design and for students to complete. These questions are highly effective in assessing factual knowledge and information recall, but they can also be used to evaluate more complex cognitive skills (Douglas et

al., 2012; Reich, 2009; Wineburg, 2004). However, a significant limitation of multiple-choice questions is that they do not reveal the more complex aspects of students' thinking and reasoning processes (Liu et al., 2023). For example, if a student selects the incorrect option 'C' on a multiple-choice item, the teacher gains no insight into the student's reasoning behind that choice. Therefore, multiple-choice assessments are less suitable for formative assessment purposes.

The Stanford History Education Group developed the History Assessments of Thinking. The assessment items are designed to elicit students' historical thinking and reasoning through short written answers that should take students no more than ten minutes to complete and are therefore less time-consuming for both students and teachers (Breakstone, 2014; Breakstone et al., 2013; Smith & Breakstone, 2015). While one drawback of these assessment items is that they are not integrated into an authentic task, they have the potential to measure complex skills in more standardised assessments. A standard format in these History Assessments of Thinking combines a multiple-choice question with a follow-up question that asks the student to explain their choice (Smith & Breakstone, 2015). The Stanford History Education Group developed design principles for their History Assessments of Thinking, informed by insights and recommendations from other scholars (Messick, 1994; Pellegrino et al., 2001) and highlighted the importance of subject and domain specificity: assessments must accurately reflect the subject content and measure subject-specific constructs. The structure of assessments that elicit complex skills must align with the targeted construct. The prompts should be designed to elicit students' complex thinking and should be as clear as possible to provide valuable insights for teachers. Furthermore, they emphasise piloting and revising as crucial iterative processes in assessment design (Breakstone, 2014; Breakstone et al., 2013; Smith & Breakstone, 2015). These principles can also serve as valuable design principles in the development process of our STAIRs.

2.3 The targeted construct in the STAIRs: Causal analysis

To measure anything effectively, it is essential first to identify the construct being assessed. For this study, we have chosen to narrow the scope of the construct. Instead of assessing proficiency in social scientific reasoning more broadly, we focus on measuring proficiency in one of its previously defined subskills (Klijnsstra et al., 2023), specifically causal analysis. This choice is primarily based on the understanding that causal analysis is a fundamental reasoning activity that underpins many other subskills, such as making predictions and drawing conclusions. As noted by Jonassen and Ionas (2008, p. 287), "Causal reasoning represents one of the most basic and important cognitive processes that underpin all higher-order activities, such as conceptual understanding and problem-solving." In the context of students' social scientific reasoning, causal analysis serves as a starting point and is, therefore, a crucial step in analysing social problems (Klijnsstra et al., 2023; Sandahl, 2015). Important subcategories in this process include reasoning with multiple causes, connecting causes and consequences, and distinguishing correlations from causation (Sandahl, 2015; Van Tubergen, 2020).

It is important to emphasise that causal analysis is closely interconnected with other subskills of students' social scientific reasoning (Klijnsstra et al., 2023). For example, causal analysis often involves attributing multiple possible explanations and underlying mechanisms, each linked to different social science paradigms, concepts, theories, or models. An additional challenge is that what is considered a social problem is socially constructed. Therefore, the causal analysis of social problems requires not only addressing multiple causal explanations but also identifying and critically

reflecting on how the problem is defined from diverse perspectives (Rosenberg et al., 1988; Sandahl, 2015, 2020; Van Tubergen, 2020).

In our previous research, we highlighted students' difficulties in reasoning (Klijnsstra et al., 2023), noting that they often confuse correlation with causation. Furthermore, students struggle to reason in the context of multiple causes and often experience difficulties in using evidence in their causal analysis. Students also tend to overestimate the role of human actions (Klijnsstra et al., 2023; Stoel et al., 2015).

Further complicating students' causal analysis in social science education is the context in which they are expected to reason: social problems. These issues are multifaceted and ill-defined, characterised by multiple causes, consequences, and potential solutions, as well as conflicting values, norms, and interests (Mills, 1959/2000; Van Tubergen, 2020). Students' preconceptions, emotions, and feelings about social problems such as youth crime, obesity, and climate change can both motivate and hinder their causal analysis (Klijnsstra et al., 2023; Sandahl, 2020; Stitzlein, 2021), adding to the complexity of the task.

This study demonstrates that the reasoning activity of causal analysis remains complex. In our previous study, we operationalised the reasoning activity causal analysis in subcategories with three proficiency levels, accompanied by descriptions of students' social scientific reasoning (see Klijnsstra et al., 2023). For example, when students confuse causes and consequences, this is operationalised as 'Beginner' (level 1) in the subcategory identifying causes. See Table 1.

Table 1. Examples of two subcategories in students' causal analysis (see Klijnsstra et al., 2023)

Subcategories	Beginner (Level 1)	Intermediate (Level 2)	Advanced (Level 3)
Identifying causes	Identifies none or only one cause of the problem	Identifies multiple causes of the problem	Identifies multiple causes of the problem and distinguishes between different types of causes, such as: <ul style="list-style-type: none"> ▪ Micro, meso, or macro scales ▪ Socioeconomic, sociocultural, or political-legal perspectives ▪ Various political-normative viewpoints ▪ Incidental versus structural causes or environmental factors
Connecting causes and consequences	Confuses causes and consequences, or does not describe how the causes contribute to the problem	Partially describes the relationships between causes and consequences in a nuanced way: <ul style="list-style-type: none"> ▪ Primarily describes relationships as deterministic and linear ▪ Pays limited attention to factors such as the direction/strength of the relationship, possible intervening causes, the distinction between causal relationships and correlations, or self-reinforcing processes 	Describes the relationships between causes and consequences in a nuanced way: <ul style="list-style-type: none"> ▪ Primarily describes relationships as probabilities ▪ Recognises that societal developments (trends) are context-dependent and subject to change ▪ Pays attention to factors such as the direction/strength of the relationship, possible intervening causes, the distinction between causal relationships and correlations, or self-reinforcing processes

In accordance with the recommendations for domain analysis and targeting subject-specific constructs (Breakstone, 2014; Messick, 1994; Mislevy et al., 2003; Pellegrino et al., 2001), the subskills and specific reasoning activities served as crucial starting points in the development of our STAIRs.

2.4 Design principles for the STAIRs

Building on insights from assessment experts (Bijsterbosch, 2018; Breakstone, 2014; Löfström et al., 2023; Messick, 1994; Mislevy et al., 2003; Pellegrino et al., 2001; Schmeiser & Welch, 2006) and consultations within the research team, four design principles guided the construction of the STAIRs. First, our assessment items must be aligned to the intended learning outcome; a specific domain construct and its subskills (Breakstone, 2014; Messick, 1994). In this case, we focused on causal analysis, utilising the rubrics that operationalised the subskills from our previous study (see Klijnstra et al., 2023).

The second design principle pertains to the contexts used: social problems. Since social scientific reasoning is inherently linked to social issues, each STAIR centres around a social problem that serves as the context for students' reasoning. Intense emotional responses might influence students' ability to conduct a causal analysis (Klijnstra et al., 2023; Sandahl, 2020; Stitzlein, 2021). When choosing contexts, designers should do their best to avoid contexts that can trigger emotional responses in a way that students underperform. For example, a standardised assessment item using domestic abuse as a context is not suitable. All students should have the opportunity to engage in causal analysis. Our STAIRs aim to minimise the need for prior knowledge of the selected social problem, providing any necessary background information as efficiently as possible.

The third design principle focuses on explicit instruction, the degree of pre-structuring, and prompts (see Breakstone, 2014). Each STAIR should include concise instructions related to the reasoning skill that is assessed. The prompt (for example, "Explain...") and subsequent instructions should provide students with clear guidance on what to do to maximise their points. This requires a careful balance in determining whether the skill being assessed includes understanding the "thinking steps" a student must take to demonstrate proficiency.

The fourth design principle addresses the difficulty and differentiation levels of the STAIRs. The STAIRs are designed to align with students' levels of social scientific reasoning and should elicit the three operationalised levels of proficiency in accordance with the conceptualisation of students' social scientific reasoning (Klijnstra et al., 2023). They should not be too difficult or too easy for the intended target group. The answer keys to score students' responses should clarify the different levels of causal analysis.

3 METHOD

3.1 Context and participants

This study was conducted in the context of social science education at the upper secondary level in the Netherlands. The assessment items (STAIRs) were developed by a design team consisting of two individuals. The first designer (Author 1) is a researcher and social science teacher educator with 15 years of experience as a social science teacher and 11 years as a teacher educator. The second designer (Author 3) has 21 years of experience as a social science teacher and is an assessment expert at the National Institute for Educational Measurement (Cito). The STAIRs were validated in

five rounds. Table 2 provides an overview of the participants involved in each validation. The STAIRs were revised after each round based on the feedback and outcomes of the validation process.

Table 2. Overview of participants per validation

Validation	Participants	Characteristics
1. Focus group	Assessment experts ($n = 7$)	<ul style="list-style-type: none"> All experts worked at the National Institute for Educational Measurement Six are social science assessment experts; one is a mathematics assessment expert Four experts have degrees in teaching social sciences Three experts have experience as social science teacher educators
2. Focus group	Social science teachers ($n = 10$)	<ul style="list-style-type: none"> Teachers were recruited from our professional network Teaching experience varied from 2 to 38 years ($M = 10,8$)
3. Think-aloud interviews/sessions	Social science students ($n = 8$)	<ul style="list-style-type: none"> Students were recruited by two teachers from our professional network Eight students from two schools participated (four students per school) Mean age: 16 years
4. Focus group	Assessment experts ($n = 3$)	<ul style="list-style-type: none"> All experts worked at the National Institute for Educational Measurement All experts have a degree in teaching upper secondary social science education Two of the three experts also participated in Validation Round 1
5. Assessment	Social science students ($n = 338$)	<ul style="list-style-type: none"> Students (social science classes) were recruited via social science teachers in our professional network 338 students from 21 classes across eight schools in both urban and rural environments Mean age: 16 years (range 15 to 18 years)

3.2 First draft of the STAIRs

Building on the previously described insights from assessment experts and related design principles, the first version of the STAIRs was designed. This initial draft included ten questions organised into four contexts, each representing social problems that students were required to reason about. The four contexts in this first version of the STAIRs were: (1) Educational level and obesity; (2) Reading skills and socioeconomic status; (3) Noise pollution from neighbours and experienced happiness; and (4) Social inequality and COVID protests. In line with the design principles, we selected these contexts because we considered them meaningful for students and have the potential to elicit students' causal analysis. In previous Dutch social science teacher education and professional development programs (Ruijs & Klijnstra, 2017, 2021), these contexts have been partly used as examples of contexts that can be used in social science education to teach complex thinking skills. Variations in the degree of pre-structuring and prompts of the questions were designed for each context and discussed in the first focus group with assessment experts. Figure 1 shows the initial STAIR 1 regarding educational level and obesity, which was discussed in the focus group with assessment experts (validation 1). As

illustrated in Figure 1, considerations about the degree of pre-structuring and different types of questions are included in the document for this first validation with assessment experts.

Figure 1. STAIR 1: Educational level and obesity

Initial STAIR 'Educational Level and Obesity'

Text 1

Obesity More Common Among People with Lower Education Levels

There is a clear correlation between lower education levels and higher rates of obesity. Among individuals with only primary education, 65% are moderately or severely overweight. In contrast, this figure drops to 35% for those with the highest levels of education. The disparity is even more pronounced when it comes to obesity, with individuals who have only primary education being more than four times as likely to be obese compared to those with a university degree.

However, this study cannot conclusively determine whether low education levels directly increase the risk of obesity, if obesity limits educational opportunities, or if other shared factors contribute to both outcomes. It is possible that all these factors play a role.

Source: RIVM

[Considerations for assessment experts version question 1A]
Version 1A is more open and less pre-structured. The question assumes that the student knows what to say about the relationship, namely: naming correlation, identifying variables, causality, correlation, direction of the relationship, etc.

Read text 1.
Question 1A [version 1A]. Make a statement about the relationship between education level and obesity according to the study in Text 1.

[Considerations for assessment experts version question 1B]
Version 1B is partially pre-structured. The question indicates which aspects a student can pay attention to but does not specify whether there is anything to say on that aspect. However, in this case, how do you deal with the scoring? Do you get points if a student mentions parts that are not applicable?]

Read text 1.
Question 1B [version 1B]. Make a statement about the relationship between education level and obesity according to the study in text 1.

You can consider the following aspects:

- the direction of the relationship
- the strength of the relationship
- the degree of causality of the relationship

[Considerations for assessment experts version question 1C]
Version 1C is highly pre-structured. Assumes familiarity with stated concepts and the simple explanations. Therefore, the version tests a smaller proportion, more questions are needed to measure the same knowledge.

Read text 1.
The three statements below are about the possible relationship between education level and obesity.

- 1 Education level and obesity correlate.
- 2 The relationship between education level and obesity is causal.
- 3 The direction of the relationship is clear.

→ For each statement, indicate whether it is true or false, or whether you cannot judge it based on text 1.

Explain your choice.

Do it as follows: copy the following and complete your answer.

Statement 1 is... (correct / incorrect / cannot be judged), because...

Statement 2 is... (correct / incorrect / cannot be judged), because...

Statement 3 is... (correct / incorrect / cannot be judged), because...

Due to the targeted construct of students' causal analysis, we utilised insights from the rubrics that operationalised three proficiency levels (see Klijnstra et al., 2023) to design the items and the answer key (scoring rubric). In the initial draft of the STAIRs, we constructed multiple instances of the same item, varying the level of pre-structuring and guidance provided to the student (see Figure 1).

3.3 Data collection

Rounds of validation

Based on the item development process of History Assessments of Thinking by Breakstone (2014) and building on recommendations from other assessment experts (e.g., Mislevy et al., 2003; Pellegrino et al., 2001; Schmeiser & Welch, 2006), we validated our STAIRs through five rounds with assessment experts, social science teachers, think-aloud interviews with students from upper secondary social science education, and by administering the items in several classes (see Table 2). The participating teachers were recruited from our own network. The project received approval from the faculty ethics committee and participants, and each validation round was audio-recorded. The five validation rounds can be categorised into three types: focus groups (Validations 1, 2, and 4), think-aloud interviews (Validation 3), and assessments in social science classes (Validation 5). Through these validation rounds, we sought to gain insights into the construct validity of the items, specifically examining the alignment between the intended reasoning levels and the reasoning elicited by the STAIRs. Additionally, we aimed to assess the difficulty of the items and their effectiveness in differentiating among students.

Focus groups (Validations 1, 2 and 4)

The previously described design principles – targeting domain-specific constructs, the appropriateness of the context (social problem), clarity of instruction, degree of pre-structuring, and difficulty and differentiation levels – served as the foundation for the development of the STAIRs, and were, therefore, the focus of our discussions in the focus groups. Furthermore, we collected additional elaborations and notes from the participants, including their insights into the STAIRs, and summarised the output of each validation round.

Before each focus group, we provided participants with the most recent version of the STAIRs along with an accompanying justification document explaining our choices. We then asked all participants to closely review and complete the items. For example, Figure 2 displays the initial version of the STAIR regarding the usefulness of facts for the city governments of Amsterdam in reducing knife violence, which was a new context we used in the focus group with social science teachers (validation 2).


Figure 2. Example of initial STAIR 2, 3 and 4 about knife violence

STAIR Knife violence [initial version of STAIRs 2, 3 and 4]

Instruction
The following resource contains information about the social problem of knife violence. Examine the source and answer the question using the knowledge you have about distinguishing (and categorizing) causes.

General information
In the Netherlands, the number of stabbings involving young people has increased in recent years. The knife violence causes a lot of social disorder. An alderman responsible for safety in Amsterdam on behalf of the VVD [right wing party] instructs his officials to draw up a plan to stop the increasing knife violence by young people in the city. Text 1 mentions three facts that may be causes of knife violence in Amsterdam.

text 1 Three facts about knife violence



fact a: Last month, Daniel (16) was stabbed in the stomach by a boy in Amsterdam East. Daniel saw who stabbed him and is willing to tell why he thinks the boy did it.

fact b: Large kitchen knives are often used in stabbings. The sale of these knives by Blokker, Hema and other shops is legal in the Netherlands, including to minors.

fact c: The youth centre in Amsterdam-South was closed in 2021. Since then, more young people have been hanging around on the streets. A sample showed that 24% of these young people had a knife in their pocket and 48% were truant at the time of the search.

Assignment
Rank the three facts from Text 1 in terms of their usefulness for officials in writing their plan.
Explain the place in your ranking for each of the facts.

Usefulness	Fact (fill in the letter)
1 most useful
2 less useful
3 least useful

Fact ranks #1. Explanation: [Initial STAIR 2]
 Fact ranks #1. Explanation: [Initial STAIR 3]
 Fact ranks #1. Explanation: [Initial STAIR 4]

In Validation 1 and 2, the assessment experts and social science teachers were asked to evaluate the extent to which they considered the STAIRs suitable for measuring the intended reasoning, how well the STAIRs matched students' reasoning levels, the extent to which they expected different levels of reasoning to be elicited, the clarity of the instructions, and how these STAIRs differed from regular social science assessment items. In each focus group, we collected participants' responses regarding the STAIRs.

Think-aloud sessions (Validation 3)

We conducted think-aloud interviews with eight social science students from two schools. These students were selected by their teachers, who were asked to choose four students, each with varying proficiency levels in social sciences. All eight think-aloud interviews were conducted at the

students' own schools. The purpose of these sessions was to identify which reasoning activities were triggered by the STAIRs and to assess the extent to which these matched the activities we intended to elicit. Students' responses during the interviews provided valuable insights into the intended guidelines of our STAIRs, including the clarity of instruction, the impact of prior knowledge related to the context, and the difficulty and differentiation levels of the items. In total, we collected 32 think-aloud responses from these eight students. Each session was audio-recorded and transcribed verbatim by Research Assistant A.

Assessments in social science classes (Validation 5)

In the final validation, 338 students in Dutch secondary social science education, spread across 21 classes in seven schools, tested the STAIRs. The objective of this fifth validation was to assess whether these items could elicit subskills of social scientific reasoning, as well as to analyse the difficulty and differentiation levels of each STAIR. The STAIRs were administered from March to May 2024. In 19 of the 21 classes, the STAIRs were introduced and conducted by one of the two research assistants or the researcher (Author 1). For practical reasons, students' social science teachers facilitated the assessments in the remaining two classes. In each class, the STAIRs were introduced to students in a similar manner: it was emphasised that, although this was an assessment setting, students would not be graded, and their teacher would not be informed of their performance. Furthermore, we emphasised that the analyses would remain anonymous and that students could opt out at any time. Students were allotted 45 minutes to complete the STAIRs.

3.4 Data analysis

The data of the focus groups (Validations 1, 2, and 4), including audio recordings, additional participants' elaborations, and notes, were summarised and prioritised by the design team. Following each validation round, and in accordance with experts' trustworthiness recommendations (Amin et al., 2020; Lincoln & Guba, 1985), the design team (Authors 1 and 3) critically discussed the validation outcomes with the other members of the research team (Authors 1, 2, and 4). The four design principles for constructing the STAIRs served as the foundation for this process. For example, we discussed feedback concerning the degree of pre-structuring (third design principle) and the relevance and usefulness of the contexts (second design principle). In this process, Authors 2 and 4 served as critical peers and collaborators, providing feedback and posing critical questions (Guba, 1981; Lincoln & Guba, 1985). After reviewing the outcomes of the validation rounds, we implemented several revisions to the STAIRs.

To analyse the think-aloud data (Validation 3), we developed a codebook to identify reasoning activities and levels of reasoning in accordance with our conceptualisation of social scientific causal reasoning (see Klijnsma et al., 2023). The research team discussed the coding scheme and made refinements as necessary. Author 1 and Research Assistant A independently scored four of the 32 student responses to evaluate the codebook's effectiveness, and differences were discussed. Following minor revisions to the codebook, Research Assistant A coded all statements made by the students during the think-aloud interviews (Validation 3). Based on the analysis of the think-aloud interviews, further refinements were made to the STAIRs.

The final version of the STAIRs was administered in social science classes (Validation 5). Figure 3 illustrates the final version of the STAIRs about knife violence as tested in social science classes (validation 5).

Figure 3. Final STAIRs 2, 3, and 4 about knife violence

General information

In Amsterdam, the number of stabbing incidents involving young people has been increasing in recent years. This rise is also evident in other major cities in the Netherlands. A city councilor in Amsterdam has instructed his officials to develop a plan to reduce knife violence among young people in the city.

text 1



Three facts about knife violence

Fact A: Last month, Daniel (16) was stabbed in the stomach by another boy in Amsterdam-Oost. Daniel noted that he had not seen any police officers in the neighborhood that day.

Fact B: Drill rap is becoming increasingly popular among young people in Amsterdam. This music genre, originating from the United Kingdom, contains lyrics that glorify and encourage knife violence.

Fact C: Research shows that young people with mild intellectual disabilities, who frequently skip school or have been expelled are more likely to carry knives. The neighborhoods in Amsterdam where knife violence is most prevalent have a relatively high number of these young people.

Evaluate how useful each of the three facts from Text 1 is for the officials developing the plan to reduce knife violence. For each fact, choose between *barely useful* and *highly useful* by crossing out the option that does not apply, and explain your choice.

Fact A is: *barely useful* / *highly useful*, because... [STAIR 2]

Fact B is: *barely useful* / *highly useful*, because... [STAIR 3]

Fact C is: *barely useful* / *highly useful*, because... [STAIR 4]

Author 1 and Research Assistants B and C coded the students' responses. Students could earn zero, one, or two points per question/STAIR. Therefore, we operationalised three levels of proficiency: beginner (zero points), intermediate (one point), and advanced (two points). Given the characteristics of our data, including students' brief written answers and the three optional proficiency levels for coding, we employed Krippendorff's alpha (KALPHA) to measure inter-coder reliability for the STAIRs. The initial stage of our coding involved a pilot phase, during which all assignments from the first 50 students were coded. The scores were compared, and any differences were discussed, resulting in refinements to the codebook. Table 3 presents an excerpt from this codebook related to the final STAIR 2 about knife violence.

Table 3. Excerpt from the codebook, building on the conceptualisation of students' social scientific reasoning (Klijnstra et al., 2023)

Reasoning activities and subcategories	Beginner	Intermediate	Advanced
Subskill STAIR 2: <ul style="list-style-type: none"> Identifying and distinguishing causes Learning to use sources and data critically: anecdotal evidence Drawing conclusions 	Based on the evidence provided, the student does not identify the type of cause, specifically distinguishing between incidental and structural causes, and fails to draw a reasoned conclusion.	Based on the evidence provided, the student identifies the type of cause, distinguishing between incidental and structural causes, but does not draw a reasoned conclusion.	Based on the evidence provided, the student identifies the type of cause, distinguishing between incidental and structural causes, and draws a reasoned conclusion.
Example of students' reasoning on STAIR: knife violence:	<p>"Fact A is very useful because they now know the police need to do more frequent checks through the neighbourhoods" (student #49)</p> <p>Alternative: "Fact A is very useful because, as an improvement, there should always be police around; otherwise, it makes no sense at all. It also allows for better monitoring" (student #50)</p>	<p>"Fact A is hardly useful as that observation is subjective, perhaps he just did not run into the police" (student #4).</p> <p>Alternative: "Fact A is hardly useful because the fact that Daniel did not see any police does not mean they were not there. It could also be that the police had an emergency that day, or they just happened not to be in the neighbourhood" (student #51)</p>	"Fact A is hardly useful because the absence of police is only a momentary observation. From Fact A, we cannot conclude that no police were present, only that Daniel did not notice them. This perspective is subjective and therefore lacks the objectivity needed to form a concrete part of an official plan to reduce knife violence" (student #17)

Subsequently, we independently coded 15% of students' responses for each assignment. Author 1 and Research Assistant B coded STAIR 1 (development of reading skills), STAIR 5 (noise pollution) and STAIR 6 (corporate fraud), while Author 1 and Research Assistant C coded STAIRs 2, 3, and 4 (all three about the context of knife violence). Depending on the specific STAIR, we required one to three rounds before achieving a reliable KALPHA (see Results section). In our quantitative analysis, we excluded any incomplete responses, including those that were illegible or only partially answered. As a result, we included 326 student responses on the STAIRs out of the total of 338 student responses in our quantitative analyses. We calculated the students' mean scores and the standard deviation. Furthermore, in line with the aim of our study and the recommendations of assessment experts (e.g., Van Berkel et al., 2017; Zegota et al., 2022), we calculated the items' degree of difficulty (p' value: 0–1.00) and the level of differentiation (RIR value and the percentages of students' scores of zero, one, and two points per STAIR). In the Result section, examples of students' answers are verbatim translations.

4 RESULTS

4.1 Validations

Validation 1: Assessment experts

A key concern of the assessment experts was the need to operationalise and substantiate the targeted domain more explicitly. They emphasised the importance of clearly articulating the specific student behaviours targeted by the STAIRs and how these behaviours corresponded with different levels of reasoning. Several experts noted that the instructions in the questions could be made more explicit. For example, the instruction “Make a statement about the relationship between educational level and obesity according to the research in Text 1” was deemed “too vague” by one expert. The instruction “Make a statement about the direction of the relationship between education level and obesity according to the study in Text 1” was preferred. However, we removed this context in the second version of the STAIRs. Partly because of the feedback on the questions, but mainly due to discussions in the research team about the context itself. More specifically, we considered the context ‘Educational level and obesity’ as potentially too sensitive: when students (or their parents) are obese, this can trigger emotional responses and can make them feel uncomfortable and hinder their reasoning. Furthermore, in retrospect, the research team questions the choice of this specific text as input for reasoning, considering recent discussions in the Netherlands regarding the terms higher and lower educated.

Furthermore, the experts emphasised the relevance of providing a more specific elaboration of students’ answers in their proficiency levels within the correction model/answer key. The assessment experts regarded the context “Social inequality and COVID protests” as too complex, noting that students’ reasoning in this context would depend too heavily on their prior knowledge and experiences. Based on this feedback, we removed this context.

Moreover, the assessment experts emphasised the importance of eliminating any ambiguity in both the instructions and the interpretation of data in the figures. This feedback, along with others, resulted in a clearer delineation of the subskills we aimed to assess. We reviewed the extent to which the required reasoning sub-steps should be explicitly provided to students. Consequently, we divided the questions into fewer sub-steps, making them less pre-structured.

Validation 2: Social science teachers

The revised version of the STAIRs included six questions, organised into four (partly new) contexts: (1) Reading skills and socioeconomic status; (2) Knife violence in Amsterdam; (3) Noise pollution from neighbours and experienced happiness; and (4) Interpreting corporate fraud. Contexts 1, 3 and 4 have functioned as examples of contexts that can be used in social science education to elicit complex thinking skills in previous Dutch social science teacher education and professional development programs (Ruijs & Klijnstra, 2017, 2021). In Validation 2, social science teachers ($N = 10$) considered the STAIRs to be innovative. Overall, teachers were positive about the initiative to assess complex skills in social sciences in this manner.

The participating social science teachers generally considered the targeted construct to be clear. They expected that the STAIRs would effectively elicit students’ social scientific reasoning. However, they offered suggestions to target specific subskills more precisely. For example, teachers recommended splitting the STAIR, which involved ranking three facts about knife violence into three

separate tasks. They argued that ranking the facts did not yield sufficient substantive information to distinguish between varying levels of social scientific reasoning, as the number and quality of arguments supporting the rankings did not necessarily indicate a higher level of reasoning in one student compared to another. Based on this feedback and subsequent discussions within the research team, we made two significant adjustments to this STAIR about knife violence. First, we modified the facts to clarify their meaning and highlight their relevance. Second, we divided the three facts into three distinct questions, each requiring students to evaluate the usability of each fact for authorities in Amsterdam in developing a plan to address knife violence. Consequently, students were asked to assess the usability of each fact individually rather than ranking them. The teachers generally viewed the contexts as useful and meaningful for students. Context 1 (Reading skills and socioeconomic status) and Context 2 (Knife violence in Amsterdam) were considered highly relevant. Context 3 (Noise pollution from neighbours and experienced happiness) was considered as relevant, although some teachers considered the title somewhat abstract. Regarding the fourth context (Interpreting corporate fraud), some teachers initially perceived it as closely related to mathematics but later emphasised its relevance for demonstrating reasoning.

Opinions were mixed on the clarity of the task instructions. One teacher felt that the instruction that students need to “reason” was still too vague. However, suggestions to provide specific thinking steps for students in their reasoning received little support in the focus group.

Validation 3: Think-aloud interviews with social science students

The contexts of the STAIRs have not changed in this third validation. Additionally, the number of questions has remained the same (four contexts and six questions: three questions regarding the context of knife violence). Based on feedback from the focus group with social science teachers, minor revisions were made to the instructions of the STAIRs. For instance, adjustments were made to the title of context 3, and data that were unnecessarily distracting (such as certain percentages in the fourth context on corporate fraud) were clarified and refined.

The think-aloud interviews conducted with social science students ($N = 8$) during Validation Round 3 revealed that students needed more time to complete the STAIRs than initially anticipated. While we initially estimated that the tasks would take 20 minutes, we later adjusted this to 40 minutes. Additionally, there was a noticeable variation in the quality of students’ responses, which aligned with the teachers’ expectations, given their selection of students at varying levels of proficiency.

The analyses of the think-aloud interviews indicated that all four STAIRs elicited reasoning activities, such as “identifies one or more causes,” “identifies the direction of the connection,” or “formulates a conclusion.” For example, one student’s response on the STAIR regarding the development of reading skills among children of parents with higher and lower levels of education was:

One can see that the [reading skills among children of] parents with high socioeconomic status actually improve during the summer holidays, likely because these parents teach their children themselves, while among those with low socioeconomic status, you can see that it declines, because the parents probably can’t teach them as much themselves.

This example reflects the reasoning of multiple students in the think-aloud interviews, as they mentioned the direction of the connection between students’ reading skills and their parents’ socioeconomic status and their formulation of a conclusion for this connection (e.g., “so they [the

parents] probably teach their children themselves”). However, most students did not explain the observed differences between the groups and what mechanism might explain this. In general, the STAIRs hardly provoked reasoning at the highest level during the think-aloud interviews.

Lastly, we discovered that the items stimulated metacognitive activities. These statements offered insight into how clear it was for the students regarding what to do. For example, students described their strategies for answering the STAIRs: “Ok (...) first it says I have to read the figure, so then I’ll do that first. And, well, after that, I study the graph. And look at the legend first.”

Validation 4: Assessment experts

The two assessment experts who participated in Validation 1 observed significant progress between the earlier and current versions. Overall, all three experts found the contexts meaningful for students and effective in eliciting student reasoning. They highlighted the innovative nature of the STAIRs and underscored the relevance of the questions and the contexts.

At a more detailed level, the experts provided additional feedback, primarily concerning the clarity of the instructions and the degree of pre-structuring and prompts. For example, they suggested that the description in the initial STAIR 2 about knife violence, “Daniel is willing to tell more about the situation,” was too vague as an introduction and might allow for excessive student interpretation. Based on this and other suggestions, we made final adjustments to the STAIRs before testing them on a larger scale in social science classrooms. For example, we have changed the introduction in STAIR 2 into: “Last month, Daniel (16) was stabbed in the stomach by another boy in Amsterdam-Oost. Daniel noted that he had not seen any police officers in the neighbourhood that day” (see Figure 3).

The four validation rounds resulted in six questions (items¹) divided across four contexts. Table 4 provides an overview of the contexts, questions, and subskills that are central to the STAIRs.

Table 4. Overview of contexts and subcategories in the final STAIRs

STAIRs	Context	Question	Subskills
STAIR 1	Development of reading skills	Referring to Figure 1, explain how and why students from low socioeconomic status (SES) backgrounds fall behind in reading skills compared to students from high SES backgrounds.	<ul style="list-style-type: none"> Comparing groups Connecting causes and consequences Interpreting data
STAIR 2	Analysing knife violence in Amsterdam	Consider how useful fact A is to local policymakers who aim to reduce knife violence. Choose from “hardly useful” and “very useful” by crossing out what does not apply, and explain your choice. Fact A: Last month, Daniel (16) was stabbed in the abdomen by a boy in Amsterdam East. Daniel said he noticed that he had not seen any police in the neighbourhood that day. “Fact A is hardly useful / very useful because...”	<ul style="list-style-type: none"> Identifying and distinguishing causes Learning to use sources and data critically: anecdotal evidence Drawing conclusions

¹ In assessment construction, “a question” is referred to as “an item.” This is because not all items are formulated as questions; they can also be constructed as instructions, hence the more generic term “item.”

STAIRs	Context	Question	Subskills
STAIR 3	Analysing knife violence in Amsterdam	Fact B: Drill rap is becoming increasingly popular among youth in Amsterdam. This music genre, which originated in Britain, features lyrics that glorify and promote knife violence. “Fact B is hardly useful / very useful because...”	<ul style="list-style-type: none"> Identifying and distinguishing causes Learning to use sources and data critically: spurious correlation Drawing conclusions
STAIR 4	Analysing knife violence in Amsterdam	Fact C: Research indicates that young people with minor mental disabilities who are truant or expelled from school are more likely to carry knives. A significant number of these individuals reside in neighbourhoods in Amsterdam where knife violence is prevalent. “Fact C is hardly useful / very useful because...”	<ul style="list-style-type: none"> Identifying and distinguishing causes Learning to use sources and data critically: relevant evidence Drawing conclusions
STAIR 5	Analysing noise pollution from neighbours	Explain that Diagram B represents the results of the CBS survey in Text 2 better than Diagram A.	<ul style="list-style-type: none"> Identifying variables Connecting causes and consequences
STAIR 6	Interpreting corporate fraud	Text 3 states that in the companies studied, “it can quickly be concluded that lower-level staff are more often guilty of fraud than higher-level staff.” However, this conclusion cannot be drawn based solely on the data from the text. Argue what additional data are needed to assess whether that conclusion remains correct.	<ul style="list-style-type: none"> Comparing groups Connecting causes and consequences Assessing the validity of a conclusion

4.2 Quantitative analyses

In this second part of the Results section, we discuss students' scores and responses on the STAIRs more quantitatively, analysing the item difficulty and discriminatory power of the STAIRs. First, we discuss the inter-rater reliability of the scoring of student answers.

Inter-rater reliability

The STAIRs were assessed in 21 social science classes (326 students in upper secondary social science education). Table 5 shows the KALPHA score for 15% of students' responses to the six assignments. Generally, scores above .667 are considered reliable (Krippendorff, 2004). The interpretation of scores depends on the complexity of the content (Hayes & Krippendorff, 2007; Krippendorff, 2004). Since this study involved coding a complex skill, the inter-rater reliability scores between .69 and .89 (see Table 5) reflected an acceptable to good level of reliability.

Table 5. Overview of scores (N = 326)

Item	Item Description	M (SD)	IRR ¹	p' ²	RIR value ³	Zero points (%)	One point (%)	Two points (%)
STAIR 1	Development of reading skills in children of parents with high and low levels of education	1.00 (.65)	.80	.50	.22	21.17	57.98	20.86
STAIR 2	Analysing knife violence in Amsterdam: anecdotal evidence	.33 (.60)	.89	.16	.13	73.93	19.33	6.75
STAIR 3	Analysing knife violence in Amsterdam: spurious correlation	.19 (.47)	.74	.10	.20	83.74	13.19	3.07
STAIR 4	Analysing knife violence in Amsterdam: relevant evidence	.77 (.56)	.87	.39	.27	29.75	63.19	7.06
STAIR 5	Analysing noise pollution from neighbours	1.06 (.76)	.69	.53	.24	26.07	42.02	31.90
STAIR 6	Interpreting corporate fraud	.50 (.74)	.79	.25	.20	64.11	21.47	14.42

Note. *M* and *SD* are used to represent the mean and standard deviation. ¹IRR = Inter-Rater Reliability based on Krippendorff's alpha, ²p' = Difficulty of the item, ³RIR value = Differentiation of the item

Level of difficulty of the STAIRs

Students could score a maximum of two points for each STAIR. As shown in Table 5, students achieved the highest mean scores on STAIR 1 about the development of reading skills ($M = 1.00$) and STAIR 5 about noise pollution ($M = 1.06$). Additionally, on average, students most frequently scored two points on STAIR 5 and STAIR 1 (see Table 5). In both STAIRs, students were required to interpret data from a figure and draw conclusions. To achieve two points for STAIR 5, students needed to describe connections between causes and consequences in a nuanced way. More specifically, a two-point score implies that a student paid attention to the direction of the relationship and distinguished between causal links and correlations. Students' two-point scores demonstrated high levels of social scientific reasoning, as they often used social scientific terms in their reasoning, even though this was not explicitly requested in the instructions. The following answer on STAIR 5 about noise pollution is exemplary:

[In Diagram] A, it is stated that the amount of perceived noise pollution is the independent variable and therefore affects the other two variables. However, the study only describes a correlation and not a causal relationship with a clear independent and dependent variable. In [Diagram] B, this correlation is shown according to the results of the study, without adding a causal relationship that is not evident from the research. (student #157)

Furthermore, we can conclude that students experienced difficulties in answering the questions related to the context involving knife violence, as reflected in their low mean scores on STAIR 2 about knife violence and anecdotal evidence ($M = .33$) and STAIR 3 about knife violence and spurious correlation ($M = .19$), which were associated with multiple reasoning flaws. The analysis reveals that in both STAIRs 2 and 3, students tended to draw conclusions prematurely based on the evidence provided. These STAIRs often demonstrated that students confused anecdotal evidence with social scientific evidence and correlations with causations. For example, in STAIR 2, one student stated: "*Fact B is very useful because he says he did not see any police that day, so officials could use more police (checks) in their plan*" (student #294). Furthermore, the highest scores (two points) were relatively scarce on STAIR 2, 3, and 4 (analysing the context of knife violence in Amsterdam).

Level of differentiation of the STAIRs

As shown in Table 5, the analysis of students' scores demonstrates RIR values above 0.2 for STAIRs 1, 4, 5, and 6. Given the relatively small sample of items and acceptable p' values, these values indicate that these items do indeed discriminate effectively. However, STAIRs 2 and 3 (both about the context of knife violence) discriminate less well, with RIR values of .13 and .18, respectively. As shown in Table 5 (see p'), on average, students achieved only 16% of the maximum score (two points) on STAIR 2, and 9% on STAIR 3.

In this test (or, more accurately, this collection of items), the number of items is very small. For this reason, a statistical analysis comparing how well items function in relation to other items is only weakly valid. However, the RIR value does inform us about the extent to which an item differentiates between proficient and less proficient students. Items 1, 4, 5, and 6 all score an RIR above 0.2, indicating that these items may effectively discriminate between students who have and have not acquired a certain level of proficiency. The low RIR scores for STAIRs 2 and 3 were likely caused by the high level of difficulty (i.e., the lack of average proficiency among students). Given that almost no students were proficient, STAIRs 2 and 3 do a poor job of distinguishing between proficiency levels.

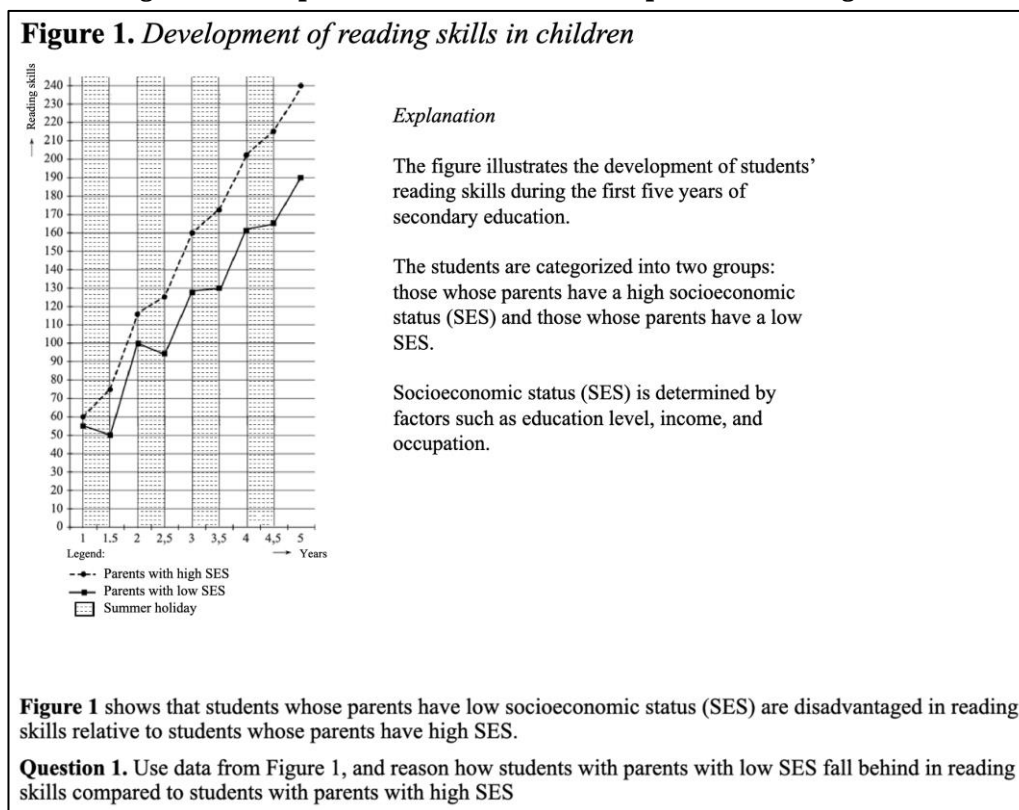
4.3 In-depth qualitative analysis of two STAIRs

In this third part of the Results section, we focus on a qualitative analysis of two specific STAIRs: 1 and 3. STAIR 1 about the development of reading skills was selected for further examination because its score distribution closely approximates a normal curve, with most values concentrated around a score of one. Additionally, STAIR 1 elicited a wide range of student reasoning across all three levels.

STAIR 3 about knife violence and identifying a spurious correlation was selected for analysis due to its low average score; students were least likely to achieve the maximum score of two points on this task. Moreover, STAIR 3 frequently revealed flaws in students' social scientific reasoning. Next, we discuss the intended reasoning steps for STAIRs 1 and 3 and the actual reasoning demonstrated by the students.

Students' reasoning related to STAIR 1

Students achieved relatively high average scores on STAIR 1 in terms of the development of reading skills. This task focuses on the subskill of connecting causes and effects using information from a given source. More specifically, in STAIR 1, students must interpret a graph to explain how students of parents with low levels of socioeconomic status (SES) are disadvantaged in reading compared to parents with high SES (see Figure 4).

Figure 4. Example STAIR 1 about the development of reading skills

To demonstrate their level of proficiency in STAIR 1 about the development of reading skills, students were required to complete two steps. First, they needed to explain that the figure illustrates how students from low SES backgrounds show less significant improvement (and sometimes even a decline) in reading skills over the summer holidays compared to their peers from high SES backgrounds. Second, students had to explain how the combination of summer holidays and low SES could contribute to a decline in reading ability. For example, students could explain that during the holidays, children are more reliant on their immediate environment; if that environment offers fewer resources, such as cultural capital (e.g., valuing reading or offering support for reading) and economic capital (e.g., access to books or financial means to purchase them), this can hinder their reading development. Table 6 provides examples of student responses to STAIR 1 about the development of reading skills for zero, one, and two points.

Table 6. Examples of students' answers on STAIR 1 about the development of reading skills

Points	Student's reasoning
2	"The graph shows that students of parents with low SES experience much less growth in their reading skills over the summer holidays compared to students of parents with high SES. This may be because parents with low SES cannot invest the same amount of money and/or resources in their children's reading development as parents with high SES. So, you see that the reading skills of students with high SES parents continue to develop gradually, while those of students with low SES parents primarily improve during the school year but do not show much improvement during the holidays" (student #258).
1	"Students of parents with high SES have access to extracurricular classes and additional books that enable them to acquire knowledge more quickly. Other children do not have access to this" (student #142)
0	"Students with low SES do not read during the vacations" (student #167).

Students who scored two points were able to identify the differences in reading skill development during the summer holidays and explain the underlying mechanism. Although the students did not use the term “cultural capital,” they effectively conveyed that families with high SES may have a greater awareness of the importance of reading, particularly regarding future education and career prospects.

Students who scored one point typically failed to explain the mechanism behind the observed differences. For example, some students, like student #142 (see Table 6), recognised that students of parents with high SES have more resources than others. However, they did not connect this observation to the idea that these resources become particularly significant when school is not in session, leaving students more dependent on their home environment during the summer holidays.

Students’ responses that scored zero points generally lacked nuances in their reasoning. Additionally, we observed that these responses often exhibited little critical reflection (see student #167 in Table 6).

Students’ reasoning related to STAIR 3

Students achieved relatively low average scores on STAIR 3 ($M = 0.19$). STAIRs 2, 3, and 4 followed the same format: in all three assignments, students received the same instruction regarding the increase in stabbing incidents in Amsterdam (see Figure 10, “General information”). However, for STAIRs 2, 3, and 4, students were presented with different factual statements. Students needed to reason to what extent each fact is useful for policymakers aiming to reduce knife crime in the city of Amsterdam. Figure 5 displays the details for STAIR 3 about knife violence.

Figure 5. Example STAIR 3 about knife violence

Text 1

General information

In Amsterdam, the number of stabbing incidents involving young people has been increasing in recent years. This increase can also be seen in other major cities in the Netherlands.

A councilor in Amsterdam orders his officials to create a plan to reduce knife violence by young people in his city.

Fact B: Drill rap is also becoming increasingly popular among Amsterdam youth. This genre of music originating from Britain contains lyrics glorifying and encouraging knife violence.

Consider how useful the fact is to the officials making the plan to reduce knife violence. Choose from “barely useful” and “very useful” by crossing out what does not apply, and explain your choice.

‘Fact B is barely useful / very useful because.....’

In STAIR 3 about knife violence (spurious correlation), students categorise causes and consequences and make reasoned inferences. More specifically, in STAIR 3, we expected students to address two components. First, they needed to identify and distinguish between correlations and causations. In addition, we expected students to determine how useful this fact is for policymakers seeking to reduce knife crime in Amsterdam. Table 7 provides examples of students’ responses on STAIR 3 about knife violence for zero, one, and two points.

Table 7. Examples of students' answers on STAIR 3 about knife violence

Points	Student's reasoning
2	"Fact B is hardly useful because it does not provide evidence of increased knife violence; it only indicates that such violence is glorified and encouraged in drill rap. Therefore, it is difficult to draw conclusions from this and create an effective plan to reduce knife violence" (student #206).
1	"Fact B is hardly useful, because it does not indicate whether stabbings related to drill rap are increasing; it only states that knife violence is glorified and encouraged" (student #169).
0	"Fact B is quite useful; you now know that there is a high probability that knife violence also occurs in Dutch drill rap groups" (student #204).

Students who scored one point often made an initial observation, noting that no evidence currently connects drill rap to stabbing incidents. However, these students did not relate this observation to the usefulness of the fact for authorities in developing a plan to reduce knife violence (see student #169 in Table 7). Overall, students who scored one point struggled to draw clear conclusions about the importance of this information for policymakers working to address knife violence. Students who scored zero points frequently misinterpreted the fact, believing it implied an increase in knife violence and attributing drill rap as a direct cause (see student #204 in Table 7). In these cases, students' reasoning tended to lack nuance, often confusing causation with correlation.

5 CONCLUSION AND DISCUSSION

Although the importance of teaching complex skills is widely recognised (Brookhart, 2010; Ercikan & Seixas, 2015; Schraw & Robinson, 2011), assessing these skills in the teaching practice remains a challenge, particularly in formative assessments (Ercikan & Seixas, 2015). This is especially relevant to the assessment of students' reasoning about social problems (Jansson, 2023; Lee et al., 2021; Sluijsman, 2013, 2014). Building on a conceptualisation of students' social scientific reasoning (Klijnstra et al., 2023), this study aimed to develop STAIRs: assessment items designed to elicit specific subskills of social scientific reasoning and measure the quality of students' reasoning through short written responses.

The key finding of this study is that the STAIRs effectively elicit subskills related to students' causal analysis of social problems. Our study demonstrates that the design principles were feasible and promising as a starting point for designing items that can elicit subskills of students' social scientific reasoning. Based on the validation of the STAIRs through focus groups with assessment experts and social science teachers, think-aloud interviews with social science students, and the analysis of the final STAIRs in social science classes (involving 338 students) using a rubric with three levels of proficiency, we can conclude that most STAIRs achieved their intended purpose: they measured subskills of social scientific reasoning. Most (but not all) items discriminated between the three proficiency levels and exhibited an acceptable level of difficulty. When students received low scores on the STAIRs, their responses often revealed reasoning flaws. These difficulties in students' reasoning aligned with findings from our previous study (see Klijnstra et al., 2023), which operationalised reasoning flaws such as excessively linear reasoning, confusing correlation with causation, and overestimating human agency. Regarding the design principles, we experimented with the degree of pre-structuring. Ultimately, our STAIRs were revised to offer fewer pre-structured sub-steps. Students' prior knowledge about the specific context may influence their reasoning, despite significant precautions to minimise the influence of context-specific prior knowledge.

For example, students with greater knowledge of social and cultural capital have an advantage in STAIR 1 about the development of reading skills.

The second design principle concerns the use of context in assessment items. Findings from this study indicate that social problems used as contextual prompts should be meaningful to students, as this can increase their motivation to engage in reasoning. When selecting appropriate contexts, designers should take into account the potential impact a given topic might have on students. Strong emotional responses, for instance, may interfere with students' ability to perform causal analysis (e.g., Sandahl, 2020; Stitzlein, 2021). One example is the context of "educational level and obesity," which was ultimately removed due to its perceived sensitivity. This is not to suggest that potentially sensitive topics should be avoided in social science education. On the contrary, social problems are inherently complex and often controversial, and such complexity should be explicitly addressed in social science education. However, this does not necessarily mean they are always suitable for use in assessment tasks. When assessing students' reasoning, particularly in formative settings, it is important to remain mindful of how contextual elements may affect students in different ways.

Social scientific reasoning is inherently knowledge-dependent: it cannot occur in isolation from what students already know about the context or about reasoning. This makes knowledge not merely beneficial, but essential for students' reasoning. Despite this, our study did not systematically address students' prior knowledge of the selected social problems, nor their familiarity with key social science concepts and terminology, such as distinctions between types of causes, or the difference between correlation and causation. Future studies could investigate the interplay between students' prior knowledge and their reasoning quality, as well as examine how different types of social problems – and the emotions or strong opinions they may provoke – influence students' reasoning processes and the types of scaffolding they might require.

The development and analysis of the STAIRs also offer insights into the types of questions that can promote social scientific reasoning. For example, STAIR 5 demonstrated that diagramming relationships helps to promote deeper reasoning. Furthermore, the use of figures and text sources may unintentionally engage other skills and influence students' reasoning. Although the STAIRs measure similar subskills of causal analysis across different question formats, a limitation of this study is the relatively small number of question types. Only one question, for example, experimented with diagrammatic representations of causal relationships. Future research that incorporates a broader range of question types would provide greater insight into the validity of specific formats. Follow-up research could experiment with similar items in different contexts and measure students' prior knowledge. In addition, lessons could be learned from other disciplines. For instance, prior research in science education has highlighted the effectiveness of model–evidence link diagrams as both instructional scaffolds and assessment tools that can promote students' complex thinking skills, such as their ability to analyse and understand complex subject-specific concepts (Lombardi et al., 2013). The use of such diagrams could further enhance our understanding of how different assessment items and formats can be employed to measure social scientific reasoning.

This study focused on developing STAIRs for a specific subskill of social scientific reasoning: causal analysis. However, social scientific reasoning encompasses other subskills, such as students' use of social scientific concepts, models, and theories, as well as their ability to adopt different perspectives and reflect on them (Klijnstra et al., 2023). As such, our findings do not address the effectiveness of STAIRs in assessing these other subskills. Future research could explore whether the

characteristics and underlying design principles of our STAIRs can be adapted to assess additional subskills related to students' reasoning about social problems.

Even though the combination of these six STAIRs could be perceived as an assessment (i.e., as a test), it is not intended to function as one. The STAIRs are not constructed as a complete test, and therefore, we make no claims about the reliability of the assessment as a test. Further research is needed to substantiate such claims.

STAIRs can be used by teachers as diagnostic tools to assess students' reasoning abilities, identify misconceptions, and track students' progress in reasoning. Student answers that scored zero points often demonstrated previously identified reasoning flaws (see Klijnsstra et al., 2023). Therefore, the STAIRs have the potential to serve as feedback tools that highlight what is lacking in students' reasoning and what they can do to achieve a higher level of reasoning. In line with the recommendations of the History Assessments of Thinking (Breakstone, 2014; Breakstone et al., 2013; Smith & Breakstone, 2015), sustained professional development and teacher training will be necessary to implement and design formative assessments like STAIRs effectively. To implement these types of items, it is essential for social science teachers, teacher educators, and assessment experts to understand the conceptualisation of the intended social scientific reasoning and to be able to distinguish subskills and proficiency levels in this area, specifically, reasoning and knowledge about how to elicit this reasoning. The design principles and underlying conceptualisation of students' social scientific reasoning can be beneficial in designing STAIRs tailored to specific contexts that are central to the social science curriculum.

Finally, the development of STAIRs fostered constructive collaboration with national assessment experts in social science education. The principles underlying the STAIRs could be applied to the development of assessments for Dutch national social science exams, paving the way for further collaboration between Dutch researchers and assessment experts.

This study constitutes an initial step in exploring how students' social scientific reasoning can be assessed. As noted by Ercikan et al. (2016), designing assessments that accurately elicit complex skills (such as social scientific reasoning) is more easily said than done. Nevertheless, this study contributes to the ongoing effort to develop new assessment methods that improve students' reasoning and social science education as a whole.

FUNDING

This research was partly funded by National Regieorgaan Onderwijsonderzoek (NRO, grant number 40.5.18540.109) and Gooise Scholen Federatie (GSF).

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

- Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M. A., Tamim, R., & Zhang, D. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of Educational Research*, 78(4), 1102–1134.
<https://doi.org/10.3102/0034654308326084>

- Amani, J., Kitta, S., Kapinga, O. S., & Mbilinyi, C. (2021). Secondary school teachers' knowledge on procedures for constructing quality classroom tests in Tanzania. *Üniversitepark Bülten*, 10(1), 40–54. <https://dx.doi.org/10.22521/unibulletin.2021.101.3>
- Amin, M. E. K., Nørgaard, L. S., Cavaco, A. M., Witry, M. J., Hillman, L., Cernasev, A., & Desselle, S. P. (2020). Establishing trustworthiness and authenticity in qualitative pharmacy research. *Research in Social and Administrative Pharmacy*, 16(10), 1472–1482. <https://doi.org/10.1016/j.sapharm.2020.02.005>
- Bijsterbosch, H. (2018). *Professional development of geography teachers with regard to summative assessment practices* [Doctoral dissertation, Utrecht University]. Utrecht University Repository. <https://dspace.library.uu.nl/handle/1874/364154>
- Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school*. National Academy Press.
- Breakstone, J. (2014). Try, try, try again: The process of designing new history assessments. *Theory & Research in Social Education*, 42(4), 453–485. <https://doi.org/10.1080/00933104.2014.965860>
- Breakstone, J., Smith, M., & Wineburg, S. (2013). Beyond the bubble in history/social studies assessments. *Phi Delta Kappan*, 94(5), 53–57. <https://doi.org/10.1177/003172171309400512>
- Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. Ascd.
- Brookhart, S. M., & Durkin, D. T. (2003). Classroom assessment, student motivation, and achievement in high school social studies classes. *Applied Measurement in Education*, 16(1), 27–54. https://psycnet.apa.org/doi/10.1207/S15324818AME1601_2
- Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, 18(1), 32–42. <https://doi.org/10.3102/0013189X018001032>
- Campbell, C. (2013). Research on teacher competency in classroom assessment. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 71–84). SAGE. <https://doi.org/10.4135/9781452218649.n5>
- College van Toetsen en Examens. (2019). *Syllabus Maatschappijwetenschappen VWO. Centraal Examen 2022. Versie 2, juli 2020* [Syllabus Social Sciences Education. Central exam. Version 2, July, 2020]. College van Toetsen en Examens.
- Cooper, A., Klinger, D. A., & McAdie, P. (2017). What do teachers need? An exploration of evidence-informed practice for classroom assessment in Ontario. *Educational Research*, 59(2), 190–208. <https://doi.org/10.1080/00131881.2017.1310392>
- Curry, K., & Smith, D. (2017). Assessment practices in social studies classrooms: Results from a longitudinal survey. *Social Studies Research and Practice*, 12(2), 168–181. <https://doi.org/10.1108/SSRP-04-2017-0015>
- Douglas, M., Wilson, J., & Ennis, S. (2012). Multiple-choice question tests: A convenient, flexible and effective learning tool? A case study. *Innovations in Education and Teaching International*, 49(2), 111–121. <https://doi.org/10.1080/14703297.2012.677596>
- Ercikan, K., & Seixas, P. C. (Eds.). (2015). *New directions in assessing historical thinking*. Routledge. <https://doi.org/10.4324/9781315779539>

- Ercikan, K., Seixas, P., Kaliski, P., & Huff, K. (2016). Assessment of history learning. In H. Braun (Ed.), *Meeting the challenges of measurement in an era of accountability* (pp. 236–267). Routledge. <http://dx.doi.org/10.4324/9780203781302-12>
- Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *ECTJ*, 29(2), 75–91. <https://doi.org/10.1007/BF02766777>
- Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1), 77–89. <https://doi.org/10.1080/19312450709336664>
- Jansson, T. (2023). Civics teachers' assessment practices in Swedish upper secondary schools. A qualitative study. *JSSE - Journal of Social Science Education*, 22(2). <https://doi.org/10.11576/jsse-5938>
- Jonassen, D.H., & Ionas, I.G. (2008). Designing effective supports for causal reasoning. *Educational Technology Research and Development*, 56, 287–308. <https://doi.org/10.1007/s11423-006-9021-6>
- Klijnstra, T., Stoel, G. L., Ruijs, G. J., Savenije, G. M., & van Boxtel, C. A. M. (2023). Toward a framework for assessing the quality of students' social scientific reasoning. *Theory & Research in Social Education*, 51(2), 173–200. <https://doi.org/10.1080/00933104.2022.2132894>
- Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). SAGE. <https://doi.org/10.4135/9781071878781>
- Lee, C. D., White, G., & Dong, D. (2021). *Educating for civic reasoning and discourse*. National Academy of Education.
- Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. In. SAGE. [http://dx.doi.org/10.1016/0147-1767\(85\)90062-8](http://dx.doi.org/10.1016/0147-1767(85)90062-8)
- Liu, Q., Wald, N., Daskon, C., & Harland, T. (2023). Multiple-choice questions (MCQs) for higher-order cognition: Perspectives of university teachers. *Innovations in Education and Teaching International*, 61(4), 802–814. <https://doi.org/10.1080/14703297.2023.2222715>
- Löfström, J., & Ouakrim-Soivio, N. (2022). Politics and ethics of civic and citizenship education curricula in Denmark, Finland, Iceland, Norway and Sweden. In R. Desjardins & S. Wiksten (Eds.), *Handbook of civic engagement and education* (pp. 182–190). Edward Elgar. <https://doi.org/10.4337/9781800376953.00025>
- Löfström, J., Rosenlund, D., & Weber, B. (2023). Assessment and national exams in social studies and social sciences. *JSSE - Journal of Social Science Education*, 22(2). <https://doi.org/10.11576/jsse-6594>
- Lombardi, D., Sibley, B., & Carroll, K. (2013). What's the alternative? Using model-evidence link diagrams to weigh alternative models in argumentation. *The Science Teacher*, 80(5), 36–41. http://dx.doi.org/10.2505/4/tst13_080_05_50
- Maddox, L. E., & Saye, J. W. (2017). Using hybrid assessments to develop civic competency in history. *The Social Studies*, 108(2), 55–71. <https://doi.org/10.1080/00377996.2017.1283288>
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13–23. <https://doi.org/10.3102/0013189X023002013>
- Mills, C. W. (2000). *The sociological imagination*. Oxford University Press. (Original work published in 1959)

- Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, 2003(1), i–29. <https://doi.org/10.1002/j.2333-8504.2003.tb01908.x>
- Moss, P. A. (2013). Validity in action: Lessons from studies of data use. *Journal of Educational Measurement*, 50(1), 91–98. <https://doi.org/10.1111/jedm.12003>
- Newmann, F., Carmichael, D., & King, M. (2016). *Authentic intellectual work*. Corwin. <https://doi.org/10.4135/9781506322308>
- Olgers, T., Meijs, L., & Dogterom, K. (2021). Een korte geschiedenis van maatschappijleer [A brief history of social science education]. In R. van den Boorn (Ed.), *Handboek vakdidactiek Maatschappijleer* (pp. 131–167). Landelijk Expertisecentrum Mens- en Maatschappijvakken.
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. National Academy Press. <https://doi.org/10.17226/10019>
- Reich, G. A. (2009). Testing historical knowledge: Standards, multiple-choice questions and student reasoning. *Theory and Research in Social Education*, 37, 325–360. <https://doi.org/10.1080/00933104.2009.10473401>
- Rosenberg, S., Ward, D., & Chilton, S. (1988). *Political reasoning and cognition: A Piagetian view*. Duke University Press. <https://doi.org/10.1215/9780822381525>
- Ruijs, G. L., & Klijnstra, T. (2017, February 2). *Denkvaardigheden bij maatschappijwetenschappen* [Thinking skills in social science education]. Paper presented at National Conference Renewed Dutch Social Science Program, Utrecht, The Netherlands.
- Ruijs, G. L., & Klijnstra, T. (2021). Hogere denkvaardigheden: denkgereedschap voor maatschappijleer [Higher thinking skills: thinking tools for social science education]. In R. van den Boorn (Ed.), *Handboek vakdidactiek Maatschappijleer* (pp. 169–205). Landelijke Expertisecentrum Mens- en Maatschappijvakken.
- Sandahl, J. (2015). Preparing for citizenship: The value of second-order concepts in social science education. *JSSE - Journal of Social Science Education*, 14(1), 18–29. <https://doi.org/10.4119/jsse-732>
- Sandahl, J. (2020). Opening up the echo chamber: Perspective taking in social science education. *Acta Didactica Norden*, 14(4), 24-sider. <https://doi.org/10.5617/adno.8350>
- Schmeiser, C. B., & Welch, C. J. (2006). Test development. *Educational measurement*, 4, 307–353. <https://doi.org/10.1016/j.pse.2014.11.006>
- Schraw, G., & Robinson, D. H. (Eds.). (2011). *Assessment of higher order thinking skills*. Information Age Publishing.
- Seixas, P., Morton, T., Colyer, J., & Fornazzari, S. (2013). *The big six: Historical thinking concepts*. Nelson Education.
- Sluismans, L. (2013). *Evaluatie van het vernieuwde examenprogramma maatschappijwetenschappen voor havo: Pilot 2010-2013* [Evaluation of the renewed examination program for social sciences for pre-university secondary education: pilot 2010-2013]. SLO. <https://www.slo.nl/@4250/evaluatie-vernieuwde-0/>
- Sluismans, L. (2014). *Evaluatie van het vernieuwde examenprogramma maatschappijwetenschappen voor vwo: pilot 2010-2014* [Evaluation of the renewed examination program for social sciences for pre-university education: pilot 2010-2014]. SLO. <https://www.slo.nl/@4248/evaluatie-vernieuwde/>

- Smith, M., & Breakstone, J. (2015). History assessments of thinking: An investigation of cognitive validity. In K. Ercikan & Peter Seixas (Eds.), *New directions in assessing historical thinking* (pp. 233–245). Routledge.
- Stitzlein, S. M. (2021). Defining and implementing civic reasoning and discourse: Philosophical and moral foundations for research and practice. In C. D. Lee, G. White, & D. Dong (Eds.), *Educating for civic reasoning and discourse* (pp. 23–52). National Academy of Education.
<https://doi.org/10.31094/2021/2>
- Stoel, G. L., van Drie, J. P., & van Boxtel, C. A. (2015). Teaching towards historical expertise: Developing a pedagogy for fostering causal reasoning in history. *Journal of Curriculum Studies*, 47(1), 49–76. <https://doi.org/10.1080/00220272.2014.968212>
- Ultee, W. C., Arts, W. A., & Flap, H. D. (2003). *Sociologie: Vragen, uitspraken, bevindingen* [Sociology: questions, statements, findings] (3rd ed.). Wolters-Noordhoff.
- Van Berkel, H., Bax, A., & Joosten-ten Brinke, D. (2017). *Toetsen in het hoger onderwijs* [Tests in higher education] (4th ed.). Bohn Stafleu van Loghum. <https://doi.org/10.1007/978-90-368-1679-3>.
- Van Boxtel, C., Hemker, A., Klijnsstra, T., & Ruijs, G. (2017). *Toetsen van denkvaardigheden en conceptuele kennis bij maatschappijwetenschappen* [Assessing thinking skills and conceptual knowledge in social science education]. Landelijk Expertisecentrum Mens- en Maatschappijvakken. <https://shorturl.at/9uVd5>
- Van Boxtel, C., & van Drie, J. (2018). Historical reasoning: Conceptualizations and educational applications. In S. A. Metzger & L. M. Harris (Eds.), *The Wiley international handbook of history teaching and learning* (pp. 149–176). John Wiley & Sons.
<https://doi.org/10.1002/9781119100812.ch6>
- Van Tubergen, F. (2020). *Introduction to sociology*. Routledge.
<https://doi.org/10.4324/9781351134958>
- William, D., & Leahy, S. (2015). *Embedding formative assessment: Practical techniques for k-12 classrooms*. Learning Sciences International.
- Wineburg, S. (2004). Crazy for history. *Journal of American History*, 90, 1401–1414.
<https://doi.org/10.2307/3660360>
- Woerdman, E. (2013). *Politiek en politicologie* [Politics and political science]. Noordhoff.
- Young, K. M., & Leinhardt, G. (1998). Writing from primary documents: A way of knowing in history. *Written communication*, 15(1), 25–68. <https://doi.org/10.1177/0741088398015001002>
- Zegota, S., Becker, T., Hagmayer, Y., & Raupach, T. (2022). Using item response theory to appraise key feature examinations for clinical reasoning. *Medical Teacher*, 44(11), 1253–1259.
<https://doi.org/10.1080/0142159X.2022.2077716>

AUTHOR BIOGRAPHIES

Thomas Klijnsstra, Ph.D. candidate at the Research Institute of Child Development and Education of the University of Amsterdam. His research focuses on students' social scientific reasoning in social science education. He works as a social science teacher trainer at the Graduate School of Child Development and Education (ILO) at the University of Amsterdam.

Geerte Savenije is an assistant professor at the Research Institute of Child Development and Education of the University of Amsterdam. Her research interests are: controversy and sensitive historical topics, the skill of historical perspective taking and learning history in museums and heritage institutions. She works as a teacher trainer at the Graduate School of Child Development and Education (ILO) at the University of Amsterdam

Chiel Huijskes is a policy advisor at the National Institute for Educational Measurement in The Netherlands. He is an assessment expert in social science education and works as a social science teacher at an upper secondary Dutch school.

Carla van Boxtel, Professor of History Education at the Research Institute of Child Development and Education of the University of Amsterdam. Her research interest includes historical reasoning, historical argumentation, the learning of historical concepts, heritage and museum education, social scientific reasoning, the potential of dialogic teaching, content-and-language integrated learning and inquiry-based learning.