European Journal of Science and Mathematics Education, 2026, **14**(3), 323-336
ISSN: 2301-251X (Online)

OPEN ACCESS

Research Article

Development of an instrument measuring cognitive load during a peer-to-peer dialogue in mathematics education

Anne Jonker [1*]
0009-0007-9730-034X

Jeroen G. Spandaw [2]
0000-0001-8793-1622

Marc J. de Vries [1]
0000-0002-1982-2157

[1] Faculty of Applied Sciences, Delft University of Technology, Delft, THE NETHERLANDS
[2] Faculty of Electrical Engineering, Mathematics and Computer Science, Delft, THE NETHERLANDS
[*] Corresponding author: a.jonker-1@tudelft.nl

Citation: Jonker, A., Spandaw, J. G., & de Vries, M. J. (2026). Development of an instrument measuring cognitive load during a peer-to-peer dialogue in mathematics education. *European Journal of Science and Mathematics Education, 14*(3), 323-336. https://doi.org/10.30935/scimath/18456

ARTICLE INFO

Received: 10 Nov 2025
Accepted: 31 Mar 2026

ABSTRACT

Peer-to-peer dialogue can enhance students' understanding of mathematics by stimulating active processing and articulation of knowledge. However, this type of interaction also places demands on working memory, which may hinder learning if cognitive load becomes excessive. To optimize classroom dialogue, it is important to distinguish between different types of cognitive load: intrinsic load (IL), extraneous load (EL), and germane load (GL). Existing self-report instruments do not account for the distinct cognitive demands associated with students' roles as listeners or explainers. This study aimed to develop and validate a questionnaire to measure IL, EL, and GL separately for both listening and explaining roles during peer-to-peer dialogue in secondary mathematics classrooms. The development process involved a literature review, analysis of existing instruments, adaptation for adolescent learners, and integration of mathematical dialogue characteristics. The resulting instrument consists of 18 items, 9 for each role. To validate the instrument, two studies were conducted using peer instruction in Dutch secondary school classes (n = 65 and n = 32; ages 15-17). Principal component analysis confirmed a three-factor structure aligned with the three types of cognitive load for both roles. The results suggest that the questionnaire is a promising tool for measuring differentiated cognitive load during classroom dialogue. It may inform instructional design aimed at balancing cognitive demand and supporting effective peer interaction in mathematics education.

Keywords: cognitive load, mathematics, secondary school, peer-to-peer dialogues

INTRODUCTION

Mathematics education is traditionally grounded in the principle of repeated practice, emphasizing the importance of sustained rehearsal. However, an alternative approach to fostering mathematical learning could lie in engaging students in peer dialogues.

Several positive effects of such a dialogue have been demonstrated among students. It helps to reveal misconceptions (Crouch & Mazur, 2001), stimulates the development of conceptual understanding (Brooks & Koretsky, 2011), and can improve student performance (Smith et al., 2009, 2011; Vickery et al., 2015). A potential concern is that this may place a high cognitive load on students (Chi, 2013). This raises the question of how much cognitive load such an instructional strategy imposes, making it worthwhile to measure this load.

In recent years, educational research has increasingly emphasized the role of cognitive resources in learning (Ouwehand et al., 2025). Learning mathematics requires effort and active use of working memory without exceeding its capacity. A widely used theory in education regarding the functioning of memory is Sweller's (2011) cognitive load theory (CLT). According to CLT, we use three primary resources when thinking: the environment, working memory, and long-term memory. The environment and long-term memory are unlimited information storage, unlike working memory, which has a very limited capacity.

When learning, our working memory is occupied by different workloads: intrinsic load (IL), extraneous load (EL), and germane load (GL). Some of the working memory's capacity is occupied by the load caused by the difficulty of the task itself, which is called IL. Working memory resources are also required by how the information is presented. This load is associated with EL. The sum of IL and EL cannot exceed the working memory's capacity if learning is to take place. To increase learning, teachers should reduce EL and optimize IL (Lovell & Caviglioli, 2020). Learning at school not only requires students' active use of their working memory, but it also requires that the working memory capacity is not exceeded. In addition to using working memory, putting effort into the acquired mathematical learning materials is also essential for learning.

Using argumentation in the classroom is a way to motivate students to make an effort and actively use their working memory. Research indicates that incorporating argumentation into the classroom is effective in promoting a deeper understanding of the material being learned (Asterhan & Schwarz, 2009). However, not all students participate productively in peer-argumentations. Participation is shaped by motivation and achievement-related attitudes as well as by prior knowledge: without explicit scaffolding, interactions often devolve into brief answer exchanges rather than explanation or justification (Berland & Reiser, 2009; Webb & Farivar, 1994). Moreover, classroom climate influences motivation and willingness to enter argumentatively rich dialogue (Ryan & Patrick, 2001). Some claim that explaining to others requires monitoring the listener's comprehension, assessing their level of understanding, and adjusting and adapting as needed. This is referred to as the transaction costs (Kirschner et al., 2009). This may exceed the available cognitive capacity and could lead to an explanation that is unclear and confusing. Neither party (explainer and listener) may benefit (Kirschner et al., 2018).

Our research approached a productive participation in a peer-to-peer dialogue from a cognitive load perspective. Having a dialogue can help students understand the material better (Tullis & Goldstone, 2020). However, it may also lead to cognitive overload if it exceeds their working memory capacity (Mohamed & Saleh, 2025). It is therefore important to understand when, for whom, and why this overload occurs. Therefore, we would like to have more insight into the cognitive load during a peer-to-peer dialogue to make it possible to optimize the teacher's guidance further. To get this insight, we need an instrument to measure cognitive load during this classroom activity. To learn how a teacher could reduce EL and increase IL in the classroom activity, the instrument must be able to make this distinction. Instruments exist to measure different types of cognitive load. However, no instrument measures the cognitive load associated with the dialogic argumentation processes. This research aims to develop an instrument that measures the different types of cognitive load during a peer-to-peer dialogue in mathematics education.

Problem Statement and Research Question

While peer-to-peer dialogue and argumentation are recognized as effective strategies to enhance mathematical understanding, they may also impose a high cognitive load on students. This cognitive burden can limit the effectiveness of such interactions, especially if it exceeds the capacity of working memory. Although instruments exist to measure different types of cognitive load, there is currently no tool designed to measure the specific cognitive demands associated with dialogic argumentation in mathematics education. To better support teachers in guiding such activities and to optimize students' learning, there is a need for an instrument that can distinguish and measure IL, ELs, and GL during peer-to-peer dialogues in the mathematics classroom. This leads to the following research question: How can the different types of cognitive load, IL, EL, and GL, be measured during peer-to-peer dialogic argumentation in mathematics education?

THEORETICAL FRAMEWORK

Cognitive Load Theory

As mentioned in the introduction, CLT's primary assumption is the limited working memory capacity (Sweller, 2011). If the available amount of working memory is exceeded, it results in an overload. This will decrease the learning effect, making the lesson less effective. An overload of working memory is a problem because it is precisely working memory that is very important in the learning process. An essential goal in learning is to process information in working memory and then transfer it to long-term memory (Sweller et al., 1998). This is how cognitive schemas are created. A schema is an organized mental structure that integrates related knowledge. As students encounter new information, they link it to what they already know, thereby building and refining schemas. Such linkages enable chunking, treating multiple related elements as a single unit in working memory, which reduces the number of separate items that must be processed and thus lowers working-memory demands (Sweller et al., 1998).

The CLT is used to draw up guidelines to improve instructions for the optimal use of working memory. Cognitive load is generally viewed as the effort someone experiences when performing a task. Originally, CLT distinguished between three types of cognitive load: IL, EL, and GL (Sweller et al., 1998). Some of the working memory load is occupied by the difficulty of the information or task itself, called 'ICL'. It increases with the number of elements the information consists of that the student must learn. The amount of IL depends on the learning goals that must be achieved; therefore, it is independent of the instructions (Sweller, 2011). So, IL is tied to the content's complexity compared with the learner's existing knowledge; EL arises when instruction is designed in ways that hinder learning. These instructions are characterized, among other things, by how the information is presented. The load required to absorb the information or by the activities students must do is called the EL. This load will always be there, but if, for example, there is a lot of redundancy in the presentation or if a lot of unclear language is used, then a relatively large amount of the available capacity will go to processing the information. There will, therefore, be a relatively large amount of EL and less capacity will be left over to process the information. This leads to the fundamental recommendation of CLT: In order to increase learning, reduce EL, and optimize IL (Sweller, 2011).

As mentioned, GL was initially introduced as the cognitive resources devoted to processes that directly facilitate learning, such as schema construction and automation. However, in a revised model of the theory, Sweller et al. (1998) abandoned GL as a distinct category. The reason for this change is that the extent to which you must make an effort to process the information depends greatly on the number of elements in the task. The GL is, therefore, strongly related to the IL. Whereas the content and its design impose IL and EL, GL is not imposed by the materials themselves. They also argue that GL is not empirically separable from IL and, conceptually, it reflects the learner's engagement with the essential complexity of the material. Therefore, these authors now recognize only two types of cognitive load: IL and EL. All cognitive effort that contributes to learning is considered by them as part of IL, while EL continues to represent avoidable and unproductive mental effort.

Despite Sweller's (2011) theoretical move to subsume GL under IL, some argue that retaining GL as a distinct construct remains valuable, particularly in the context of instructional design and measurement. For example, de Jong (2010) and Leppink et al. (2013) have argued that GL captures the learner's intentional effort to engage in deep processing, which is not fully explained by intrinsic task complexity. From this perspective, GL reflects the strategic allocation of cognitive resources to meaningful learning processes, such as elaboration, self-explanation, or abstraction, which can vary even when IL remains constant.

Moreover, measurement studies have shown that learners report distinguishable experiences of effort associated with understanding versus effort spent on dealing with poor instructional design (Kalyuga, 2011; Leppink et al., 2014). Research (DeLeeuw & Mayer, 2008; Klepsch et al., 2017; Leppink, 2016) has shown that it is possible to distinguish among these three different types of cognitive load. In practical terms, distinguishing GL may help instructors optimize their design by evaluating not just whether students can handle task complexity, but also whether they are actively engaging in productive cognitive strategies.

In sum, while some researchers have merged GL and IL conceptually, there are still compelling reasons, especially from an instructional and empirical standpoint, to measure GL separately. Doing so provides insights into how learners engage with the learning process, not merely what the task demands.

Cognitive load while listening or explaining

In this context, listening is defined as the active process of not only perceiving auditory information but also trying to understand its meaning and intent. Effective listening requires constant interaction between auditory processing and cognitive interpretation, whereby students integrate incoming information with prior knowledge. This process involves executive functions such as attention regulation, planning, monitoring, reasoning, and evaluating. Peer dialogue in mathematics alternates between listening and explaining, and each role engages working memory and executive control in distinct ways consistent with CLT.

Listening entails decoding, interpreting, and integrating new information with existing knowledge (Mayer, 2002). When listening, IL increases with element interactivity and gaps in prior knowledge, thereby taxing limited-capacity resources (Sweller et al., 1998). By contrast, EL during listening is driven by presentation-induced inefficiencies such as disorganized discourse, unclear articulation, split attention between speech and visuals, redundancy, and suboptimal pacing, all of which consume capacity without aiding schema construction (Kalyuga et al., 1999). Listeners incur GL when they deliberately allocate effort to reorganize, refine, and integrate the message with existing schemas through inference-making, coherence monitoring, and conceptual re-mapping (Paas et al., 2003).

In addition to these core sources of cognitive load, several modulating factors may influence listening performance. Sustained attention can wane over time, particularly in monotonous or extended listening situations, increasing the risk of cognitive disengagement and elevating EL. Emotional and social dynamics, such as anxiety about peer evaluation or frustration due to comprehension difficulties, may consume cognitive resources that would otherwise support processing (Beilock, 2008). Furthermore, the perceived relevance of the content affects GL: students are more likely to invest mental effort when they view the topic as meaningful or interesting (Paas et al., 2003).

In peer learning, two qualitatively different kinds of student explanations often emerge. Peer-explaining refers to explanations that actively build understanding: the explainer monitors what is known and unknown, links ideas to prior knowledge, draws inferences, and articulates the principles that make a solution work. These explanations reorganize information for an audience and aim to advance shared conceptual clarity. In contrast, peer-telling is largely transmissive: students provide answers, recite facts, or list steps with little effort to justify, connect, or evaluate them. Whereas peer-explaining engages metacognitive and integrative processes that tend to deepen the explainer's own learning, peer-telling primarily delivers information without substantially transforming it (Roscoe & Chi, 2007).

For this article, we use explaining as an umbrella term that includes both peer-explaining and peer-telling. In other words, explaining refers to student-to-student exchange of information, talking with a classmate about the task, whether the talk primarily builds understanding (linking ideas, justifying, monitoring) or mainly conveys information (stating answers, facts, or steps). Our analyses, therefore, encompass the full range of peer verbalizations that occur during collaborative work, from meaning-making dialogue to more straightforward information delivery.

When students explain, IL comes from the built-in complexity of what is being explained, the number of elements that must be coordinated, the logical order of steps, the use of multiple representations, and the management of subgoals, while keeping the explanation propositionally coherent (Sweller, 2010). EL during explaining often stems from coordinating two tasks at once (thinking and talking), monitoring a peer's understanding, interpreting verbal and nonverbal cues, and adapting the message on the fly, as well as evaluation anxiety; these demands pull attention away from task-relevant processing and can reduce fluency (Eysenck et al., 2007).

Finally, explaining can increase GL in a beneficial way: generating, organizing, and tailoring an explanation reveals knowledge gaps, triggers elaboration and reorganization, and helps consolidate schemas, mechanisms shown in research on self-explanation and learning-by-teaching (Chi et al., 1994; Fiorella & Mayer, 2013).

Measuring cognitive load

Cognitive load can be assessed using several established methods, including physiological measures, dual-task techniques, and subjective self-report instruments (Sweller, 2011). Physiological measures, such as heart rate variability or pupil dilation, offer objective indicators of cognitive effort. However, these methods are primarily suited for laboratory contexts and are often impractical for classroom-based research, particularly with adolescents. Dual-task methods involve assigning participants a secondary task in addition to their primary learning activity. For example, students may be asked to respond to auditory or visual cues while solving a math problem. Performance on the secondary task is inversely related to the cognitive load of the primary task. While informative, this approach can interfere with natural classroom interactions. Self-report questionnaires remain the most practical method for measuring cognitive load in educational settings. One widely used instrument is the mental effort scale developed by Paas et al. (2003), in which learners rate their perceived mental effort throughout a learning task. Despite their subjective nature, such instruments have demonstrated surprising accuracy in reflecting cognitive load.

To differentiate between IL, EL, and GL, self-report instruments are currently the only option. This study builds upon existing questionnaires but extends them to include role-specific items for peer-to-peer interactions, tailored to the cognitive profile of secondary school students.

Role- and Domain-Specific Determinants of Cognitive Load in Peer-to-Peer Mathematics Dialogues

Beyond the general principles of CLT, the load experienced during peer-to-peer interaction is shaped by the learner's role, explainer or listener, with distinct demands on attention, monitoring, and formulation; mathematics-specific didactical features, such as sequential reasoning and shifts between symbolic and verbal representations; and learner characteristics typical of adolescence, including executive functions, metacognition, and social-evaluative pressure. These contextual factors modulate the profile of IL, EL, and GL, influencing the development of the role-sensitive measurement approach within mathematics dialogues.

Cognitive development in adolescents

We developed a questionnaire for secondary school students. Adolescents (ages 15-17) are in a critical phase of cognitive and executive function development, including abstract reasoning, attentional control, and metacognitive regulation. These developmental characteristics may heighten sensitivity to cognitive load during instructional interactions. For instance, students may encounter difficulties in structuring explanations (GL) or filtering distractions (EL), which must be considered when designing age-appropriate questionnaire items.

In addition, developmental differences in working memory capacity and processing speed may influence how many problem elements students can manage simultaneously. Adolescents also vary significantly in their use of metacognitive strategies, such as self-monitoring or planning, which can impact both their effectiveness in learning and their experience of cognitive load. Moreover, emotional regulation in peer contexts is still developing; fear of negative peer evaluation or social comparison may increase EL, particularly in dialogic activities.

Language

Research on questionnaire development with adolescents shows that item wording should be kept clear and concrete, as cognitive interview studies have identified unclear wording, undefined technical terms, vague expressions, and difficult vocabulary as common sources of misunderstanding in adolescent respondents (Park & Kwon, 2021).

Mathematical dialogue

Mathematics dialogue differs from discourse in other subjects in several ways. It typically involves a single correct answer, although multiple strategies may lead to that answer. The subject matter is often abstract and requires sequential reasoning, where missing a step can impede comprehension. The pressure to avoid mistakes may intensify EL.

Table 1. Comparison of words used in existing questionnaires

Load type	Leppink et al. (2013)		Klepsch et al. (2017)		Ayres (2006) (IL), Cierniak et al. (2009) (EL), and Salomon (1984) (GL)	
	Expression that is used for making a statement	Subject on which a statement is requested	Expression that is used for making a statement	Subject on which a statement is requested	Expression that is used for making a statement	Subject on which a statement is requested
IL	Complex	Topic/formulas/ concept/ definition	Complex/ keep in mind simultaneously	Task/number of elements	Easy/difficult	The lecture
EL	Unclear/ ineffective	Language/lecture	Exhausting/ inconvenient/ difficult	Find information/ design of the task/important information	Easy/difficult to learn	The lecture
GL	Enhance understanding	The activity	Effort/ understanding/ comprehension	Details/ everything/ the task	Concentrate	The task

Furthermore, explaining mathematical reasoning often requires switching between symbolic (e.g., formulas) and verbal representations. This translation process can be added to IL, especially for students who are less fluent in expressing abstract ideas in natural language. Also, since each step in a mathematical procedure builds upon previous steps, comprehension is often cumulatively dependent, a disruption in understanding at one point can compromise the entire explanation. Finally, the emphasis on arriving at the 'correct' solution may reduce students' willingness to experiment or reflect, potentially reducing GL while increasing EL due to performance pressure. These domain-specific features influence both the nature and magnitude of cognitive load during peer interactions in mathematics.

Existing Questionnaires

Several questionnaires aim to distinguish between IL, EL, and GL, for example, those developed by Leppink (2016) and Klepsch et al. (2017). These instruments were translated into Dutch and administered in upper secondary mathematics classes. Following a peer discussion on a mathematical topic, students completed one of the translated questionnaires.

However, our analysis revealed that these instruments did not yield valid or reliable results in the context of peer-to-peer interaction in secondary education. One likely explanation is that these questionnaires were originally designed for settings involving teacher-led instruction and individual problem-solving, rather than collaborative dialogue. Additionally, the age of the respondents may have made a difference.

In peer dialogue, students alternate between the roles of explainer and listener, each of which involves qualitatively different cognitive demands. For example, if an explainer articulates their reasoning clearly, the listener may experience reduced EL. Conversely, the explainer may experience increased EL due to the effort required to monitor and adapt to the listener's understanding.

Existing instruments do not account for these role-dependent dynamics. A suitable questionnaire should therefore distinguish not only between the three types of cognitive load, but also between the listener and explainer roles within dialogic settings. Prior to this paper, no known studies or instruments have addressed this specific need.

QUESTIONNAIRE DEVELOPMENT METHODOLOGY

To formulate the items for the new questionnaire, a multi-step development process was undertaken. Existing cognitive load questionnaires were systematically analyzed to identify how each instrument conceptualized IL, EL, and GL, and which instructional aspects they targeted (e.g., complexity of the material, clarity of language). These findings are summarized in **Table 1**. **Table 1** illustrates how existing questionnaires express load types and link them to instructional elements. It includes comparisons between the instruments of Leppink et al. (2013) and Klepsch et al. (2017) (both of which distinguish IL, EL, and GL), and single-load instruments developed by Ayres (2006) (IL), Cierniak et al. (2009) (EL), and Salomon (1984) (GL).

Table 2. Aspects to take into account when measuring cognitive load

	Listening	Explaining	Peer-to-Peer	Mathematics	Age
IL	Interpreting Integrating new information with existing knowledge Complexity of the content Insufficient background knowledge	Complexity of the content Interacting elements Logical sequencing Maintaining coherence Insufficient background knowledge		Requires sequential reasoning Requires often switching between symbolic and verbal representations	Less developed listening skills Less developed metacognitive strategies
EL	Disorganized discourse Unclear articulations Lack of structure Overuse of jargon Redundancy Lack of visual aids Suboptimal pacing	Multitasking requirements: monitoring listeners understanding comprehension verbal/non-verbal cues Adjusting strategies Fear of judgement	Same jargon Not equally motivated Fear of negative pear evaluation Social comparison	One correct answer (fear of being wrong)	Difficult to structure thoughts Filtering distractions
GL	Challenging pre-conceptions Refining mental models	Challenging pre-conceptions Refining mental models			Make an effort to structure the explanation

A key motivation for designing a new instrument was the recognition that existing questionnaires do not differentiate between the cognitive demands associated with the roles of listener and explainer. Since these roles impose different types and intensities of load, the new questionnaire was explicitly constructed to address this gap.

To inform this role-based distinction, a literature review was conducted on the cognitive skills involved in listening and explaining, respectively. In addition, a second review focused on best practices for designing questionnaires suited to adolescent respondents, considering their developmental characteristics and linguistic preferences. Finally, research on mathematical dialogue was examined, particularly regarding how the correctness of answers (as opposed to opinion-based dialogue) may influence perceived cognitive load.

This is summarized in **Table 2**. **Table 2** shows the five aspects (listening, explaining, peer-to-peer, mathematics and age) that influence the development of the questionnaire, plotted against the three forms of cognitive load. The first two columns describe the two different roles that a student can have, namely listening and explaining. The last three columns represent the three aspects that can specifically influence the cognitive load for this target group. For each combination, it is indicated what this form of cognitive load can cause. For example, in **Table 2**, you can see that the use of unclear language can cause an EL while listening. Another thing you can read in **Table 2** is that, specifically for this target group, the fear of peer assessment can cause an EL when conducting a dialogue.

Building on the mapping of constructs in **Table 1** and the role- and context-specific analysis in **Table 2**, **Table 3** specifies the targets and wording choices for our new questionnaire. **Table 3** operates cognitive load in peer dialogue by crossing role (listening vs. explaining) with load type (IL, EL, and GL). For each cell, we first identify the topic (what the item is about) and then the expression (how we ask about it) to minimize construct overlap and to ensure role-specific relevance. IL topics focus on content complexity and element interactivity; EL topics capture presentation, structure, and social–evaluative factors that are not inherent to the content; GL topics index the learner's deliberate investment of effort in organizing, integrating, and refining understanding. Because some activities in peer dialogue (e.g., listening while planning a response) inherently involve multiple concurrent elements, we treat "multitasking" as a compositional case and split it into listening, processing, responding, and doing these simultaneously, so that IL (element interactivity), EL (distraction/format), and GL (productive effort) can be distinguished at item level. **Table 3** thus provides role-sensitive, load-specific item subjects that will form the base for Dutch student-friendly questions for upper-secondary mathematics students.

For each combination, we formulated three questions. For example, for the combination of listening with GL and the item 'understanding', we formulated the question 'I put effort into understanding my classmate's explanation'. In this way, a questionnaire of 18 questions was created. 9 questions per role, which the student can be (listener or explainer), and for each role, 3 questions per type of load. The questions are formulated in Dutch and are included in **Table 4**. This also includes a translation of the questions.

Table 3. Subjects for the new questionnaire

Load type	Listening		Explaining	
	Expression that is used for making a statement	Subject on which a statement is requested	Expression that is used for making a statement	Subject on which a statement is requested
IL	Difficult	Listen Understanding Process Multitasking	Difficult	Convert thoughts into words The exercise convey to someone multitasking
EL	Difficult	Follow (non-linear, disorganized) Listen & think respond irrelevant information	Difficult	Talk to classmate Monitor adapt/adjust ignore distractions/stimuli
GL	Effort	Listening Understanding Processing Challenging pre-conceptions	Effort	Convert thoughts into words How to explain How to convey

Table 4. Questions in the questionnaire

	Listening	Explaining
IL1	Ik vond het moeilijk om te luisteren naar mijn klasgenoot [I found it difficult to listen to my classmate].	Ik vond het moeilijk om de stappen die ik in mijn hoofd had gezet onder woorden te brengen [I found it difficult to put into words the steps I had worked out in my head].
IL2	Ik vond het moeilijk om de uitleg van mijn klasgenoot te begrijpen [I found it difficult to understand my classmate's explanation].	Ik vond het moeilijk om te bedenken hoe ik de opgave moest uitleggen [I found it difficult to figure out how to explain the exercise].
IL3	Ik vond het moeilijk om de uitleg van mijn klasgenoot te verwerken [I found it difficult to process my classmate's explanation].	Ik vond het moeilijk om mijn uitleg goed over te brengen op mijn klasgenoot [It was hard for me to explain things in a way my classmate could understand].
EL1	Ik vond het moeilijk om de uitleg van mijn klasgenoot te volgen [I found it difficult to follow my classmate's explanation].	Ik vond het moeilijk om tegen een klasgenoot te moeten praten [I found it difficult to have to talk to a classmate].
EL2	Ik vond het moeilijk om tegelijkertijd te luisteren en te bedenken of ik het eens was met dat wat mijn klasgenoot uitlegde [I found it difficult to listen and, at the same time, think about whether I agreed with what my classmate was explaining].	Ik vond het moeilijk om in de gaten te houden of mijn klasgenoot mijn uitleg wel begreep [I found it difficult to monitor whether my classmate understood my explanation].
EL3	Ik vond het moeilijk om tegelijkertijd te luisteren en te bedenken hoe ik hierop moest reageren [I found it difficult to listen and, at the same time, think about how I should respond].	Ik vond het moeilijk om mijn uitleg aan te passen aan het begrip van mijn klasgenoot [I found it difficult to adjust my explanation to my classmate's level of understanding].
GL1	Ik heb moeite gedaan om te luisteren naar mijn klasgenoot [I made an effort to listening to my classmate].	Ik heb moeite gedaan om de stappen die ik in mijn hoofd had gezet onder woorden te brengen [I made an effort to put into words the steps I had worked out in my head].
GL2	Ik heb moeite gedaan om de uitleg van mijn klasgenoot te begrijpen [I made an effort into understanding my classmate's explanation].	Ik heb moeite gedaan om te bedenken hoe ik de opgave moest uitleggen [I made an effort to figure out how to explain the exercise].
GL3	Ik heb moeite gedaan om de uitleg van mijn klasgenoot te verwerken [I made an effort to process my classmate's explanation].	Ik heb moeite gedaan om mijn uitleg goed over te brengen op mijn klasgenoot [I made an effort to explain things clearly to my classmate].

Validation

To validate the newly formulated questions we prepared a mathematics lesson about probability for a Dutch secondary school class in two different academic streams (senior general secondary education and pre-university, age range 15-17). To provoke a peer-to-peer dialogue, we used peer instruction. Peer instruction is a class activity in which the teacher shows a multiple-choice question on the white screen. It is important that the wrong answers are known preconceptions. The students think silently and show their solution by holding a colored flashcard. Students with different solutions discuss their solutions with each other. This is the part with the peer-to-peer dialogue. The dialogue took about 10 minutes. Finally, the teacher goes through the different solutions. Directly after the peer instruction, 65 students completed the

Table 5. Mean (standard deviation), skewness, kurtosis, and component loadings for listening in study 1

Item	Mean (standard deviation)	Skewness	Kurtosis	Component loading		
				C1	C2	C3
First component						
Item 1	1.7059 (.90098)	1.312	1.938	**.901**	.007	.171
Item 2	2.3529 (1.18023)	.253	-1.156	**.739**	.070	-.213
Item 3	2.1961 (.98020)	.649	.129	**.813**	.207	.042
Item 4	1.8235 (1.03355)	1.161	.693	**.683**	-.262	-.247
Second component						
Item 5	1.7059 (.94433)	1.227	.551	.055	**.860**	-.048
Item 6	1.5294 (.73083)	1.334	1.483	-.007	**.846**	-.058
Third component						
Item 7	2.0200 (1.20357)	.838	-.364	-.126	-.038	**-.946**
Item 8	2.2800 (1.24605)	.890	.034	.200	.015	**-.796**
Item 9	2.2000 (1.17803)	.609	-.556	.056	.248	**-.783**

questionnaire with a 5-point Likert scale. The questionnaire was introduced with a short oral explanation. Students were told that the questions referred only to the peer-to-peer dialogue and not the teacher's explanation afterwards. Furthermore, we emphasized that the questionnaire was anonymous; if, for example, a given explanation was unclear, the researchers couldn't find out who had given this explanation. All the students filled in the form and handed it in directly afterwards.

A principal component analysis (PCA) in SPSS was conducted on the 9 listening items and on the 9 explaining items with oblique rotation (direct Oblimin). The Kaiser-Meyer-Olkin (KMO) measure verified that the data were likely suitable for factor analysis, KMO (.92) (listening) and KMO (.783) (explaining), and all KMO values for individual items were greater than .63 (listening) and .74 (explaining). The scree plot revealed in both analyses three factors with eigenvalues greater than 1. In combination, these factors explained over 70% (listening) and 75% (explaining) of the variance.

Although the size of the data set is small, the results of the KMO (.770) and the Bartlett's test of sphericity ($p < 0.01$) are sufficient to perform a factor analysis. We used Oblimin with Kaiser normalization as the rotation method because a correlation between the components is expected. As previously discussed, a correlation is expected between IL and GL. When the learning material is more complex or involves a higher number of interacting elements, learners are required to invest more mental effort to comprehend and solve the task. In other words, for less experienced learners, a high IL is likely to be accompanied by a high GL, as more cognitive resources are needed to process and make sense of the material. Conversely, more experienced learners dealing with less complex content are expected to exhibit both lower IL and lower GL. However, when EL is high, it consumes available working memory capacity, leaving fewer resources for meaningful processing. This may result in a reduced GL, even when the task itself requires deep engagement.

Reliability analysis for the three components revealed Cronbach's alpha values for listening (.838, .740, .861). For explaining these values showed (.881, .761, .819). The inter-item correlation shows that there is a mutual correlation, but the items are not unique. This gives no reason to revise the items. The post-extraction commonalities had no low communality (< 0.30), so there are no items that explain little of the communal variance and could potentially be removed.

RESULTS

The results are shown in **Table 5** (listening) and **Table 6** (explaining).

Analysis

The aim of this study was to develop questionnaire items that differentiate between the three types of cognitive load, IL, EL, and GL, within the context of peer-to-peer dialogue in secondary mathematics education. Our hypothesis posited that items 1 to 3 measure IL, items 4 to 6 measure EL, and items 7 to 9 measure GL in both the listening and explaining versions of the questionnaire.

Table 6. Mean (standard deviation), skewness, kurtosis, and component loadings for explaining in study 1

Item	Mean (standard deviation)	Skewness	Kurtosis	Component loading		
				C1	C2	C3
First component						
Item 1	1.4118 (.94184)	2.508	5.656	**.420**	.212	**-.534**
Item 2	1.6327 (.69803)	.652	-.695	-.005	-.082	**-.941**
Item 3	1.7959 (.91241)	.769	-.544	.015	.174	**-.867**
Second component						
Item 4	2.0200 (1.20357)	.838	-.364	-.098	**.773**	-.006
Item 5	2.2800 (1.24605)	.890	.034	-.042	**.774**	-.082
Item 6	2.2000 (1.17803)	.609	-.556	.222	**.791**	.056
Third component						
Item 7	2.1458 (1.09135)	.621	-.464	**.761**	.047	-.178
Item 8	2.1702 (1.02828)	.646	-.094	**.856**	-.122	-.176
Item 9	2.3617 (1.16890)	.691	-.226	**.874**	.065	.177

A PCA was conducted on the nine items of the listening questionnaire using oblique rotation (direct Oblimin). The KMO measure (KMO = .658) indicated that the sample was adequately suited for factor analysis, with all individual KMO values above 0.53. The scree plot revealed a clear three-factor structure with eigenvalues greater than one, which jointly accounted for over 75% of the variance in the data. This supports the assumption that the three theoretical load types are reflected in distinct components.

The first component captured items 1 to 3, consistent with IL, as these items addressed the perceived difficulty of processing the explanation. Item 4 was originally hypothesized to reflect EL, yet exhibited substantial cross-loadings between the first and third components. This may indicate ambiguity in how students interpret the phrasing of the item. Specifically, "difficulty following the explanation" may be perceived as either task-related or structurally-related, blurring the boundary between IL and EL.

The second component included items 5 and 6, which describe distractions and irrelevant information, aligning well with EL. While this is a theoretically coherent factor, a two-item component is potentially fragile and could benefit from an additional item to increase reliability and measurement precision.

Items 7 to 9 loaded strongly on the third component, associated with GL. These items assessed effortful engagement and attempts to refine understanding, which are core to the construct. The high internal consistency (Cronbach's alpha = .903) supports the robustness of this subscale.

Reliability analyses confirmed acceptable internal consistency for all three subscales: .763 (IL), .724 (EL), and .903 (GL). This lends further support to the instrument's psychometric soundness, especially given the modest sample size.

Second Study

In a second study, the listening questions that showed cross-loadings or minimal contribution to a factor were revisited. Question 4 received a more precise formulation that explicitly referenced the structure or clarity of the explanation (I found it difficult to recognize structure in my classmate's explanation). In a similar design, these questions were revalidated. 102 students completed the questionnaire. The results are shown in **Table 7**.

A PCA was conducted on the 9 listening items and items with oblique rotation (direct Oblimin). The KMO measure verified the data were likely suitable for factor analysis, KMO = .781 and all KMO values for individual items were greater than .70 . The scree plot revealed in both analysis three factors with eigenvalues greater than 1. In combination, these factors explained over 82% (listening).

Reliability analysis for the three components revealed Cronbach's alpha values for listening (.864, .874, .906). Question 4 now clusters with EL.

DISCUSSION

This study set out to design and validate a role-sensitive questionnaire that distinguishes IL, EL, and GL as experienced during peer-to-peer dialogue in secondary mathematics. Across two classroom

Table 7. Mean (standard deviation), skewness, kurtosis, and component loadings for listening in study 2

Item	Mean (standard deviation)	Skewness	Kurtosis	Component loading		
				C1	C2	C3
First component						
Item 1	2.6000 (.93203)	-.107	-.484	-.056	.011	**.868**
Item 2	2.6078 (.94562)	.936	-.841	-.020	.088	**.898**
Item 3	2.6078 (.99660)	.614	.183	.198	-.094	**.818**
Second component						
Item 4	2.5800 (.95537)	.299	-.108	**.833**	.009	.091
Item 5	2.5248 (1.01582)	.311	-.641	**.865**	.096	.033
Item 6	2.5300 (1.00960)	.337	-.370	**.933**	-.037	-.022
Third component						
Item 7	2.8515 (1.18647)	-.036	-.837	-.174	**-.898**	-.070
Item 8	3.0100 (1.07774)	-.317	-.639	.064	**-.919**	.007
Item 9	3.0200 (1.08227)	-.382	-.563	.192	**-.917**	-.026

implementations, the analyses support the intended construct: principal component analyses yielded a coherent three-factor structure for both listening and explaining, indicating that students can meaningfully differentiate the cognitive demands of dialogic interaction by load type and by role. The instrument proved feasible to administer immediately after authentic peer instruction activities, reinforcing its practical utility for classroom research.

At the same time, several item-level findings nuance this positive picture. In the listening scale, the item tapping "difficulty following the explanation" displayed notable cross-loadings between IL and EL. This pattern is theoretically plausible, students may attribute difficulty either to the inherent complexity of the content (IL) or to the way the peer's explanation is structured and delivered (EL), but it signals wording ambiguity that should be reduced in future iterations. Indeed, a rephrased version clustered more clearly with EL in the second study, though some residual overlap with IL remained. Additionally, the EL (listening) factor in the first study relied on two strong indicators; expanding this subscale with an additional item would likely improve precision and reliability.

Internal consistencies for the three subscales were acceptable to high for both roles, a notable finding given the modest sample sizes. Together with the stable three-factor solution, these results provide initial evidence for construct validity and reliability of a brief, 18-item, role-specific measure deployable in naturalistic lessons. Nevertheless, the small convenience samples and the observed cross-loadings limit generalizability; confirmatory factor analysis (CFA), test-retest reliability, and measurement invariance across educational tracks and grade levels form clear next steps.

Although CLT was developed primarily with direct instruction in mind, the present findings indicate that CLT's core distinctions transfer to collaborative, dialogic settings. In fact, a role-differentiated lens appears essential: explaining places additional demands on monitoring a peer's understanding and adapting one's message (raising EL risk), whereas listening emphasizes integrating sequential reasoning and connecting new information to prior knowledge (raising IL and, when effortful, GL). This alignment between empirical patterns and theorized role demands strengthens the case for analyzing dialogue through a CLT framework.

The results also contribute to the ongoing debate about GL. While more recent formulations of CLT have subsumed GL under IL, our three-factor solution and reliable GL subscales suggest that learners experience and can report a qualitatively distinct, productive form of effort during dialogue, particularly when they are actively reorganizing understanding or making their thinking explicit. Retaining GL as a measured construct thus appears pragmatically valuable for diagnosing and designing instruction, even if theoretical boundaries remain contested.

The questionnaire is relevant for both scientific research and classroom practice. In research, it provides a more differentiated measure of cognitive load during peer-to-peer dialogue by distinguishing between load experienced during listening and load experienced during explaining. This makes it possible to examine more precisely how task design, scaffolding, instructional conditions, and learner characteristics affect different aspects of cognitive load in collaborative settings. In doing so, the instrument may contribute to a more detailed understanding of the processes through which peer dialogue supports learning. In educational

practice, the questionnaire can be used to evaluate peer-learning activities and to identify whether students primarily experience difficulty in understanding their peers or in expressing their own ideas. These insights can inform targeted adaptations to instruction, such as providing additional structure, language support, or guidance during collaborative tasks.

The studies used small, context-specific samples, so larger and more diverse cohorts are needed, along with stronger tests such as CFA and test-retest reliability. Several items also warrant refinement: wording should make "difficulty following" clearly about structure and clarity (EL) rather than content complexity (IL); at least one additional indicator should strengthen the EL (listening) factor; and new items could address social pressure, keeping track of multi-step reasoning, staying focused while listening, and organizing one's thoughts while explaining. Future work should examine measurement invariance across tracks and grades, explore use in other subjects, and test whether changes in scores reflect real learning gains or the effects of specific teaching adjustments.

CONCLUSION

In sum, this work offers an actionable, classroom-ready instrument that differentiates IL, EL, and GL for both listening and explaining during peer dialogue in mathematics. The emerging evidence of construct validity and reliability, together with clear pathways for refinement, positions the tool as a useful bridge between CLT-informed design and the realities of dialogic, student-centered instruction.

Author contributions: AJ; conceptualization, data curation, investigation, formal analysis, and writing-original draft. **JGS** Conceptualisation, methodology, formal analysing and writing review & editing. **MJdV:** supervision and writing-review & editing

Funding: This article was funded by the Doctoral Grant for Teachers–PvL 2021-1 from the Dutch Research Council with grant number 023.017.050.

Ethics declaration: All participants provided informed consent prior to participation. They were informed about the aim of the study, the nature of their participation, and their right to withdraw at any time without consequences. All collected data were anonymized and handled confidentially to ensure participants' privacy. The data were used solely for research purposes and will not be retained longer than strictly necessary. Ethical approval for this study was obtained from the Human Research Ethics Committee of Delft University of Technology.

AI statement: During the preparation of this work, the author(s) used an artificial intelligence tool to improve readability and language. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication. No artificial intelligence tools were used at any stage of the research process, including data collection, analysis, or interpretation.

Declaration of interest: The authors declared no competing interests.

Data availability: Data generated and analyzed during this study are available from the authors on request.

REFERENCES

Asterhan, C. S. C., & Schwarz, B. B. (2009). Argumentation and explanation in conceptual change: Indications from protocol analyses of peer-to-peer dialog. *Cognitive Science, 33*(3), 374-400. https://doi.org/10.1111/j.1551-6709.2009.01017.x

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction, 16*(5), 389-400. https://doi.org/10.1016/j.learninstruc.2006.09.001

Beilock, S. L. (2008). Math performance in stressful situations. *Current Directions in Psychological Science, 17*(5), 339-343. https://doi.org/10.1111/j.1467-8721.2008.00602.x

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education, 93*(1), 26-55. https://doi.org/10.1002/sce.20286

Brooks, B. J., & Koretsky, M. D. (2011). The influence of group discussion on students' responses and confidence during peer instruction. *Journal of Chemical Education, 88*(11), 1477-1484. https://doi.org/10.1021/ed101066x

Chi, M. T. (2013). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology* (pp. 161-238). Routledge.

Chi, M. T., De Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*(3), 439-477. https://doi.org/10.1207/s15516709cog1803_3

Cierniak, G., Scheiter, K., & Gerjets, P. (2009). Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? *Computers in Human Behavior, 25*(2), 315-324. https://doi.org/10.1016/j.chb.2008.12.020

Crouch, C. H., & Mazur, E. (2001). Peer Instruction: Ten years of experience and results. *American Journal of Physics, 69*(9), 970-977. https://doi.org/10.1119/1.1374249

De Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science, 38*(2), 105-134. https://doi.org/10.1007/s11251-009-9110-0

DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology, 100*(1), 223-234. https://doi.org/10.1037/0022-0663.100.1.223

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion, 7*(2), 336-353. https://doi.org/10.1037/1528-3542.7.2.336

Fiorella, L., & Mayer, R. E. (2013). The relative benefits of learning by teaching and teaching expectancy. *Contemporary Educational Psychology, 38*(4), 281-288. https://doi.org/10.1016/j.cedpsych.2013.06.001

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology, 13*(4), 351-371. https://doi.org/10.1002/(SICI)1099-0720(199908)13:4<351::AID-ACP589>3.0.CO;2-6

Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need? *Educational Psychology Review, 23*(1), 1-19. https://doi.org/10.1007/s10648-010-9150-7

Kirschner, F., Paas, F., & Kirschner, P. A. (2009). Individual and group-based learning from complex cognitive tasks: Effects on retention and transfer efficiency. *Computers in Human Behavior, 25*(2), 306-314. https://doi.org/10.1016/j.chb.2008.12.008

Kirschner, P. A., Sweller, J., Kirschner, F., & Zambrano R, J. (2018). From cognitive load theory to collaborative cognitive load theory. *International Journal of Computer-Supported Collaborative Learning, 13*(2), 213-233. https://doi.org/10.1007/s11412-018-9277-y

Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology, 8*, Article 1997. https://doi.org/10.3389/fpsyg.2017.01997

Leppink, J., Paas, F., Van Gog, T., van Der Vleuten, C. P., & Van Merrienboer, J. J. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction, 30*, 32-42. https://doi.org/10.1016/j.learninstruc.2013.12.001

Leppink, J. (2016). Cognitive load measures mainly have meaning when they are combined with learning outcome measures. *Medical Education, 50*(9), Article 979. https://doi.org/10.1111/medu.13126

Leppink, J., Paas, F., Van der Vleuten, C. P. M., Van Gog, T., & Van Merrinboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods, 45*(4), 1058-1072. https://doi.org/10.3758/s13428-013-0334-1

Lovell, O., & Caviglioli, O. (2020). *Sweller's cognitive load theory in action*. John Catt Educational Ltd.

Mayer, R. E. (2002). Multimedia learning. In B. H. Ross (Ed.), *Psychology of learning and motivation* (pp. 85-139). Elsevier. https://doi.org/10.1016/S0079-7421(02)80005-6

Mohamed, N., & Saleh, S. (2025). Brainwaves and higher-order thinking: An EEG study of cognitive engagement in mathematics tasks. *International Electronic Journal of Mathematics Education, 20*(4), Article em0852. https://doi.org/10.29333/iejme/16889

Ouwehand, K., Lespiau, F., Tricot, A., & Paas, F. (2025). Cognitive load theory: Emerging trends and innovations. *Education Sciences, 15*(4), Article 458. https://doi.org/10.3390/educsci15040458

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*(1), 1-4. https://doi.org/10.1207/S15326985EP3801_1

Park, E., & Kwon, M. (2021). Testing the digital health literacy instrument for adolescents: Cognitive interviews. *Journal of Medical Internet Research, 23*(3), Article e17856. https://doi.org/10.2196/17856

Roscoe, R. D., & Chi, M. T. (2007). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. *Review of Educational Research, 77*(4), 534-574. https://doi.org/10.3102/0034654307309920

Ryan, A. M., & Patrick, H. (2001). The classroom social environment and changes in adolescents' motivation and engagement during middle school. *American Educational Research Journal, 38*(2), 437-460. https://doi.org/10.3102/00028312038002437

Salomon, G. (1984). Television is easy and print is tough–The differential investment of mental effort in learning as a function of perceptions and attributions. *Journal of Educational Psychology, 76*(4), 647-658. https://doi.org/10.1037/0022-0663.76.4.647

Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C. E., Knight, J. K., Guild, N., & Su, T. T. (2009). Why peer discussion improves student performance on in-class concept questions. *Developmental Biology, 331*(2), 416. https://doi.org/10.1016/j.ydbio.2009.05.104

Smith, M. K., Wood, W. B., Krauter, K., & Knight, J. K. (2011). Combining peer discussion with instructor explanation increases student learning from in-class concept questions. *CBE-Life Sciences Education, 10*(1), 55-63. https://doi.org/10.1187/cbe.10-08-0101

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review, 22*(2), 123-138. https://doi.org/10.1007/s10648-010-9128-5

Sweller, J. (2011). Cognitive load theory. In J. P. Mestre, & B. H. Ross (Eds.), *The psychology of learning and motivation* (pp. 37-76). Elsevier. https://doi.org/10.1016/B978-0-12-387691-1.00002-8

Sweller, J., van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251-296. https://doi.org/10.1023/A:1022193728205

Tullis, J. G., & Goldstone, R. L. (2020). Why does peer instruction benefit student learning? *Cognitive Research: Principles and Implications, 5*(1), Article 15. https://doi.org/10.1186/s41235-020-00218-5

Vickrey, T., Rosploch, K., Rahmanian, R., Pilarz, M., & Stains, M. (2015). Research-based implementation of peer instruction: A literature review. *CBE-Life Sciences Education*, *14*(1). https://doi.org/10.1187/cbe.14-11-0198

Webb, N. M., & Farivar, S. (1994). Promoting helping behavior in cooperative small groups in middle school mathematics. *American Educational Research Journal, 31*(2), 369-395. https://doi.org/10.3102/00028312031002369