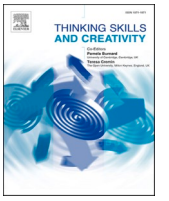
Thinking Skills and Creativity 48 (2023) 101291

Contents lists available at ScienceDirect

Thinking Skills and Creativity

journal homepage: www.elsevier.com/locate/tsc



Teaching talk for thinking: The efficacy of a peer talk teaching program for improving group thinking

Liru Hu *, Gaowei Chen, Jiajun Wu

Faculty of Education, The University of Hong Kong, Hong Kong, China

ARTICLE INFO

Keywords:
Group thinking
Productive peer talk
Talk move
Dialogic collaborative problem-solving

ABSTRACT

When a group of individuals collaboratively consider about something, they are engaged in what is referred to as group thinking. Effective group thinking is typically characterized by high-quality peer talk in which students may or may not spontaneously engage. Despite routinely providing students with multiple opportunities for collaboration, teachers rarely provide explicit instructions on effective peer talk. This study developed a teaching-talk-for-thinking program for fourth-grade students and examined its efficacy by comparing an intervention class (both learning talking and practicing collaboration) with a comparison class (only practicing collaboration). The results showed significant differences between the two classes in collaborative discourse, but not in group thinking test scores. Successful groups from the two classes improved their group thinking in contrasting manners; as a dialogue-orientated group culture emerged in the intervention class, but not in the comparison class. The study demonstrated the impact of the program in promoting the interdependence of individual members which is difficult to spontaneously learn through practice. However, further research is needed to determine if additional teaching of productive peer talk can better improve group thinking skills than accumulation of collaboration experience alone.

1. Introduction

When a group of people collaborate, they consider each other's thinking and interact to monitor and influence each other's knowledge, emotions, and actions. Mercer (2002, 2013) has described this social form of human thinking as "interthinking", as it requires a shared mental context. Stahl also observed the existence of such group-level thinking. Stahl (2016) found that "the utterances of people in dialogue produce a cognitive stream that is not attributable to either speaker individually but is a group process that only makes sense as such" (p. 363). He assigned the name "group cognition" to such an emergent cognitive stream that is only explainable at a group level. Woolley et al. (2010) even found a statistical factor called "collective intelligence" that could explain and predict group performance over a wide variety of tasks.

Based on prior research, the present study views group thinking as a collective cognitive process that occurs when a group of individuals collaboratively consider something. The quality of group thinking indicates how students work through the problem solving steps. Therefore, examining the quality of group thinking can help us better understand the process and outcomes of collaborative problem solving. The quality of group thinking may inform their solution quality and individual social and cognitive gains obtained from collaboration.

* Corresponding author at: 525 Meng Wah Complex, Pokfulam Road, Hong Kong, China.
E-mail address: liruhu@connect.hku.hk (L. Hu).

https://doi.org/10.1016/j.tsc.2023.101291
Received 27 October 2022; Received in revised form 21 March 2023; Accepted 5 April 2023
Available online 7 April 2023
1871-1871/

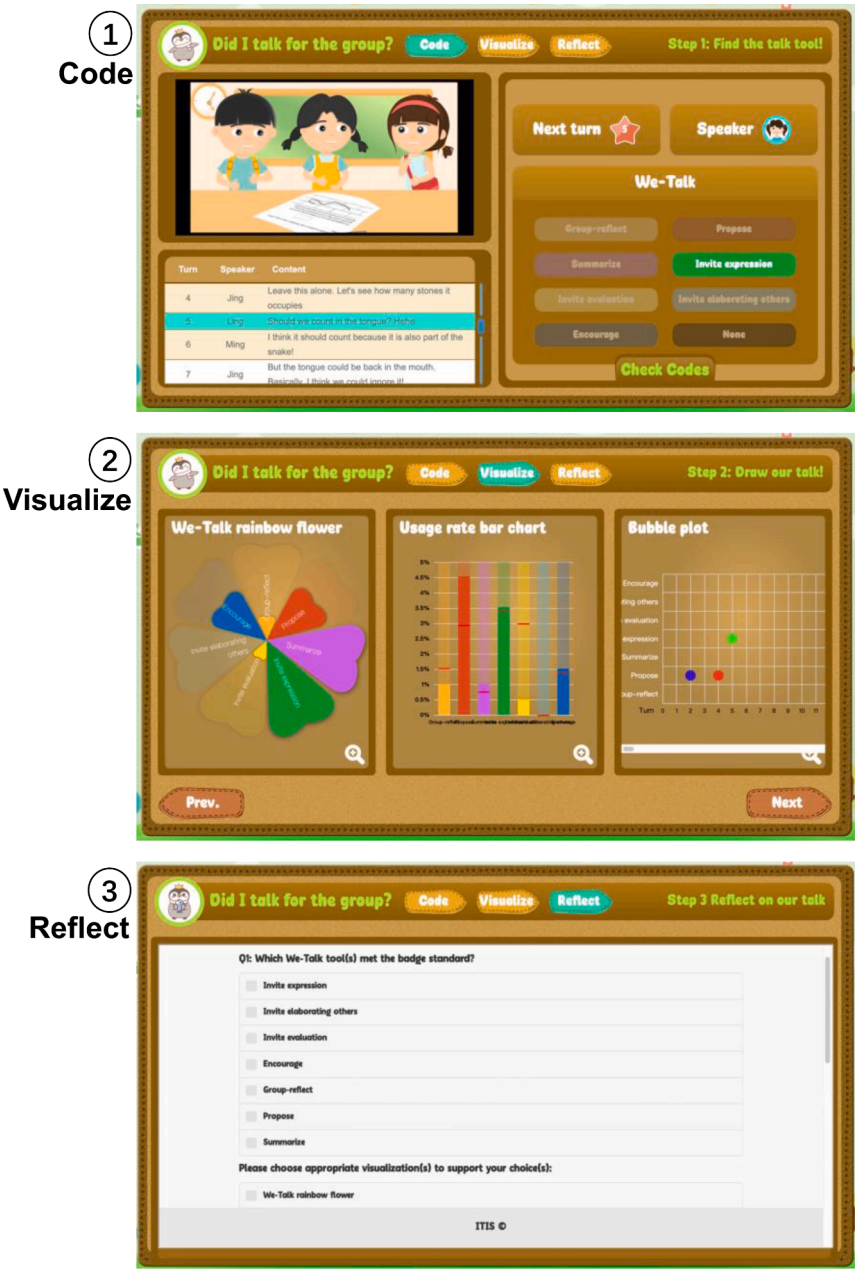


Fig. 1. Screenshots of the three-step affordances of iTalk–iSee in supporting the learning of We-Talk tools. Step 1: Code. Students access to synchronized collaboration video and transcripts and identify any We-Talk tools turn by turn. They can also check their codes by comparing to the reference codes provided by the teacher. Step 2: Visualize. Students access to three different types of visualizations concerning the usage of We-Talk tools. The We-Talk rainbow flower shows to what extent the usage rates of various tools met the standards set by the teacher. The usage rate bar chart illustrates the usage rates of various tools with additional short red lines indicating the average levels of the class. The bubble plot shows when each tool was used by a particular member. Step 3: Reflect. Some structured questions guide students to reflect on their strengths and weaknesses in using We-Talk tools and think of specific remedial actions (Hu et al., 2022).

Group thinking can be limited by emotional conflicts, poor communication skills, cultural differences, and asymmetries in social status (Chen et al., 2012a,b). Woolley et al. (2010) found that collective intelligence was correlated not only with a group's composition features (i.e., percentage of female students and individual's social sensitivity) but also its interaction patterns (i.e., participation equality).

High-quality group thinking is usually characterized by high-quality peer talk which is essential to achieve the social and cognitive benefits of collaborative problem solving in education (Gillies, 2019; King, 2008). However, students usually suffer from a set of talk-related issues when collaborate, such as social dominance or isolation (Borge & Carroll, 2014; Woolley et al., 2010), social loafing

(Simms & Nichols, 2014), rudeness and conflicts (Chiu & Khoo, 2003; Xie et al., 2013), and the unconstructiveness of simply disputational or cumulative talk (Mercer, 2002).

Many studies have sought to extract detailed talk moves (e.g., "what do you think about…," "why…," "I think…," "I agree/disagree with…," "can you say more about…," and "do you agree or disagree with….") that typify productive collaborative discourse (e.g., Gillies, 2019; Webb et al., 2014). These peer talk moves aim to perform or elicit high-level cognitive activities (e.g., explaining, questioning, arguing, monitoring, evaluating, reflecting, and summarizing) and promote social interdependence. They can therefore promote high-quality group thinking. Many studies have attempted to provide students with these productive peer talk moves to support collaborative interaction and promote group thinking and learning (e.g., Gillies, 2019; King, 1997; Noroozi, Teasley et al., 2013; Webb et al., 2014). In practice, however, teachers seldom explicitly teach students how to engage in collaboration, despite creating multiple collaboration opportunities for their students. There is also a lack of technical support for teachers in many talk intervention programs (e.g., Littleton & Mercer, 2013; Topping & Trickey, 2013).

The present study built on previous efforts and developed a technology-advanced teaching-talk-for-thinking program to improve group thinking in collaborative problem-solving. This program adopts iTalk–iSee, a participatory visual learning analytical tool (Hu et al., 2022), to teach students talk in an academically productive approach. iTalk–iSee supports young learners' development of skills in using productive peer talk moves through its three-step affordances: code → visualize → reflect (see Fig. 1). It engages students to identify productive peer talk moves in their discussion, enables them to intuit the features of their collaboration through various visual representations of collaborative discourse, and encourages them to have productive and reflective discussions about their interactions (Hu et al., 2022).

It is a common practice for teachers to arrange group activities in their classrooms. A group may spontaneously collaborate better through repeated collaboration practice, which could confound the results of a study. Therefore, in this study, the efficacy of the teaching-talk-for-thinking program is examined by comparing an intervention class to a comparison class where students receive an equivalent amount of collaboration practice. Specifically, the study addresses two research questions:

Research question 1. Was group thinking improved in the intervention class relative to the comparison class?

Research question 2. How did the intervention help to improve group thinking?

2. Theoretical background

2.1. Dialogue and thinking

Dialogue is usually viewed as a mediator for promoting thinking (Gillies, 2019; Resnick et al., 2010). Talk moves in dialogue can perform local social and cognitive functions, which not only facilitates individual thinking but also stimulates and extends group thinking. For example, the "reflect on oneself" talk move can deepen one's own thinking as they engage in metacognitive activities; the "press for reasoning" talk move can stimulate others to deepen their thinking by eliciting justifications; and the "add on others" talk move can extend group thinking through the co-construction of knowledge. Additionally, there is neuroscience evidence supporting the relationship between language skills and cognitive skills (Bialystok et al., 2005). The socio-cultural theory views language as an essential cultural and psychological tool that links individual and group thinking in a reciprocal relationship and enables the spiral intellectual development of human society (Vass & Littleton, 2010; Vygotsky, 1978; Wertsch, 1991). As Alexander (2005, p. 2) notes, "talk vitally mediates the cognitive and cultural spaces" and "language not only manifests thinking but also structures it." Individuals gain cultural understandings from elders-led and culturally-framed ways of communication. Such cultural intersubjectivity further transforms individuals' thinking by improving their metacognitive and self-regulation skills (Mercer, 2013).

The present study is situated in the theory of dialogism, which views dialogue as the space where truth or knowledge emerges (Bakhtin, 1981). Dialogism theory understands dialogue in a broader sense than the notion of "conversation" in daily life (Dressman, 2004). Dialogue is not a superficial discursive form but the interanimation of different co-equal voices through which meaning naturally emerges, akin to an electric spark triggered by two hooked terminals (Trausan-Matu et al., 2021). Dialogue can also occur within one's mind, where individuals generate dialogic responses when they utilize existing knowledge to make sense of something they hear or read (Mercer, 2013). Rather than viewing dialogue as a mediator in promoting thinking, we interpret dialogism as viewing dialogue as a form of thinking in itself. Socio-cultural theory assumes a predetermined outside and aims to erase the difference by, for example, teaching a novice to think like an expert (Wegerif, 2013). Conversely, dialogism attempts to create various voices that are equally valued and considered, thus making a difference. Such dialogic interaction denies the existence of correct views or predetermined ends ahead of dialogic interaction (Kolikant & Pollack, 2020; Matusov et al., 2019). Thinking naturally unfolds and flows amongst individuals, and learning and knowledge emerge.

2.2. Dialogic collaborative problem solving

We applied dialogism to the context of collaborative problem solving and proposed the term "dialogic collaborative problem solving", which refers to a complex dynamic process in which two or more consciousnesses, with equal rights and each with its own world, combine but are not merged in the unity of solving a shared problem (Hu & Chen, 2021, 2022). We have also identified three essential goals, or "talk virtues", for effective dialogic collaborative problem-solving: equity, open-mindedness, and convergence. Goal 1 involves establishing and maintaining team organization, while goal 2 requires establishing and maintaining shared understandings. Goal 2 can be further divided into two aspects: goal 2–1 involves elaborating and justifying one's own perspectives, while goal 2–2 involves engaging with the perspectives of others. Goal 1 and 2 align with the spirit of dialogic interaction, which emphasizes equity

Table 1
Talk tools in iTalk–iSee for use in dialogic collaborative problem-solving (Hu et al., 2022).

Goals/Talk virtues	Type of talk tools	Talk tool	Example
Equity	We-Talk-Equity	1. Invite expression	What's your opinion?
		2. Invite evaluation	Do you agree?
		3. Invite building on others	Do you have any further thoughts about his position?
		4. Encourage	Right or wrong, just say whatever you think!
Convergence	We-Talk-Convergence	1. Summarize	I find that there are some patterns in our solutions.
		2. Group-reflect	Our understandings of the handout are still different!
		3. Propose	Let's move on to the next question.
Open-mindedness	I-Talk	1. Share information	I have done similar tasks before…
		2. Express new idea	My opinion is…
		3. Build on oneself	I have a further thought about my previous opinion.
		4. Explain oneself	Here's my reasoning against your claim…
		5. Self-reflect	I am not sure about what I am saying.
	You-Talk	1. Press for elaboration	What is a new example of …?
		2. Press for explanation	Explain how your proposed solution would work.
		3. Revoice	Do you mean…?
		4. Build on others	Here's a further thought offered in the spirit of your position …
		5. Evaluate	The limitation of your claim is …
		6. Compare	What you said was the same as what I mean.

Table 2
Demographic information of the students in the intervention and comparison classes.

	Intervention (n = 59)		Comparison (n = 59)		t/χ^2	p
	M	SD	M	SD		
Age	9.59	0.50	9.61	0.56	$t(116) = -0.18$	0.86
Gender	0.59	0.50	0.61	0.49	$\chi^2(1) = 0.14$	0.71
Mother's education level[a]	2.56	1.01	2.77	1.40	$t(85) = -0.80$	0.43
Father's education level[a]	2.77	1.04	2.85	1.20	$t(85) = -0.31$	0.76
Recent mathematics grade[b]	84.76	19.65	83.90	22.37	$t(115) = 0.22$	0.82
Recent Chinese grade[b]	93.07	18.29	96.38	18.62	$t(115) = -0.97$	0.33

Note.
[a] Education level: 1: primary school or below; 2: middle school; 3: high school or technical high school; 4: junior college; 5: undergraduate; 6: graduate or above.
[b] The maximum score was 120.

amongst voices and open-mindedness of individuals (Bakhtin, 1929/1984). Goal 3 reflects an essential feature of collaborative problem-solving, which involves the convergence of individual efforts towards an optimal joint solution (Baker et al., 2020).

Productive peer talk moves encourage learners to verbalize their viewpoints and underlying methods of reasoning, and to verify, evaluate, and build upon the contributions of their peers. Based on a synthesis of studies on the efficacy of productive peer talk moves (Hu & Chen, 2023), there is a set of validated talk moves in by the existing literature. We further conceptualize these moves as four types of tools to help students achieve the goals of dialogic collaborative problem-solving (see Table 1). These talk tools are categorized through a first- or second-person perspective (i.e., I-Talk, You-Talk, We-Talk) to create a conversational tone that is more likely to benefit student learning than a formal style (Mayer, 2014).

3. Method

3.1. Participants

The participants were recruited from a low-ranking primary school in a third-tier city in China. The school principal volunteered this school to participate in the study. The school had six large and academically comparable fourth-grade classes, with approximately 60 students per class.

A power analysis was conducted using G*Power (version 3.1.9.6) (Faul et al., 2007) to determine the required sample size. Our meta-analysis on the effectiveness of productive peer talk moves (Hu & Chen, 2023) assumed a large effect size to detect anticipated differences in collaborative discourse and solution quality. Specifically, a power calculation based on an independent t-test (setting parameters alpha = 0.5, power = 0.8, tails = two, effect size d = 0.95, allocation ratio = 1) returned a total sample size of 38 (i.e., 19 in each group).

Therefore, we randomly chose two classes from this pool of six academically comparable classes and randomly assigned them as the intervention class (n = 59) and the comparison class (n = 59), respectively. One student in the comparison class participated only in the

latter phase of the program due to long-term sick leave. There were 20 groups in each class. Most groups are triads with only one or two dyads in each class. There were no significant differences between the two classes in terms of students' demographic information or prior academic performance (see Table 2). The mathematics teacher of the comparison class had approximately 10 years of teaching experience, whereas the teacher of the intervention class was younger and had approximately 3 years of teaching experience. Both teachers were female and in charge of their class.

3.2. Settings

The study was implemented as an independent semester-long elective course entitled "Mathematics Dialogue and Thinking". Each lesson was delivered in a classroom with a video-recording system and lasted for approximately 50 mins. The video-recording system had three separate cameras for the teacher, students, and electronic whiteboard, respectively.

As the above-described classroom video-recording system did not capture the discussion of each group, a group video-recording device was designed for this study. This device contained two mobile phones: one that performed video recording and one that performed backup audio recording. Due to the limited space in the classroom, a fish-eye lens was used in the front-facing camera of the mobile phone to ensure that it captured all the group members. All recorded videos and audios were automatically uploaded to a back-end server that was connected to the iTalk–iSee server.

3.3. Materials

The development of tasks in this study involved four steps. Firstly, we selected a pool of tasks from authoritative sources, based on criteria such as having right solutions and easily evaluable solution steps, being process-open, focusing on reasoning skills, and being suitable for group discussion. These sources included Po Leung Kuk Primary Mathematics World Contest (Zhu & Sun, 2018), Australian Mathematics Competition (Australian Math Trust, n.d.), International Mathematics Assessment for Schools (IMAS, n.d.), Trends in International Mathematics and Science Study survey conducted in 2015 (TIMSS & PIRLS International Study Center, 2015), and the Junior Mathematical Olympiad (Database of Mathematical Olympiad, n.d.). Secondly, we consulted experienced mathematics teachers in the school to help select a smaller pool of tasks that were suitable for fourth-grade students, and further refined their details. Thirdly, we conducted a pilot study in a class not participating in the study to test item difficulty and suitability for group discussion. Finally, we finalized a set of tasks for this study and organized them in a sequence based on their type and item difficulty (see Table A1 in the Appendix). These tasks are more similar to those found in students' textbooks, but differ from the visual puzzles used in the GTM.

3.4. Measures

We utilized two measures, the Group Thinking Measure (GTM) developed by Wegerif et al. (2017) and the Group Thinking Sustainability (GTS) developed by Hu (2021), to assess group thinking in our study. Additionally, we employed psychometric scales to measure relevant individual characteristics such as mathematics self-concept, mathematics learning enjoyment, social anxiety, and perspective-taking ability. To ensure the validity of these scales, we conducted a pilot study with a sample of 283 students. To assess internal consistency, we used Cronbach's alpha to examine the extent to which subsets of items produce similar scores (Cronbach, 1951). Although a desirable alpha value is typically set above 0.70, it is important to consider the specific context when interpreting this value (Taber, 2018).

We also evaluated the construct validity of the scales using exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). In EFA, items were removed if their factor loadings were less than 0.40, cross-loadings were greater than 0.30 (Hinton et al., 2014), or the difference between the cross-loading and the highest factor loading was less than 0.20 (Bedford, 1997). In CFA, we employed several model fit indices recommended by prior research, including the chi-square with corresponding degrees of freedom and level of significance, the Root Mean Square Error of Approximation (RMSEA) with corresponding 90% confidence intervals, the Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI), and the Standardized Root Mean Square Residual (SRMR) (Worthington & Whittaker, 2006). Typically, RMSEA values below 0.08 and SRMR values below 0.1 indicate acceptable model fit (Worthington & Whittaker, 2006), while CFI and TLI values above 0.9 (Awang, 2015), RMSEA values close to 0.06, and SRMR values close to 0.08 indicate satisfactory fit (Hu & Bentler, 1999).

3.4.1. Group thinking measure (GTM)

The GTM test builds on similar measures that have been used in many studies (Mercer et al., 1999; Rojas-Drummod et al., 2003; Wegerif et al., 1999; Woolley et al., 2010), and has been validated as effective for use with students and for studying group thinking in mathematics (Fujita et al., 2019; Wegerif et al., 2017). The GTM consists of 30 visual puzzles that are akin to Raven's Standard Progressive Matrices. Each puzzle contains an array of eight shapes and a blank space, and students are required to choose one shape from eight options that would fit in the blank space in a way that corresponds to the underlying pattern of the puzzle (refer to Fig. A1 in the Appendix). All of the puzzles are allocated into two equally difficult 15-item test sets, which are labelled A and B for assignment to either individuals or groups. The puzzles in each set become progressively more difficult. In addition, the order in which the test sets are completed does not significantly influence the test scores (Wegerif et al., 2017).

The GTM test is designed to measure the ability of educational programs to improve individual and group thinking (Fujita et al., 2019; Wegerif et al., 2017). This is important, as an improvement in a group thinking test score may be due to an improvement in the individual thinking of group members rather than in the group's collective thinking. Thus, the GTM measures the effectiveness of

collaboration by examining whether a group outperforms the maximum score of its members in identical tests. The grouping value is then determined by subtracting the highest member score from the group score. Following this, the groups are classified into three categories based on the grouping value (Fujita et al., 2019; Wegerif et al., 2017). These categories are: a value-added group, with a score that is more than one standard deviation higher than the highest member score; a value-detracting group, with a score more than one standard deviation lower than the highest member score; and a value-neutral group, with a score between the above two types. Additionally, it has been proposed that a qualitative analysis of the collaboration process based on video-recorded group discussions serves as a complementary measure to the quantitative GTM (Wegerif et al., 2017).

3.4.2. Group thinking sustainability (GTS)

GTS is a measure of group thinking that assesses the group's ability to maintain high-level cognitive thinking during group discussion. It provides a comprehensive view of group thinking by considering various patterns of interaction and group outcomes. GTS is characterized by a three-level hierarchy, including Reciprocal Group Thinking Sustainability (RGTS), Productive Group Thinking Sustainability (PGTS), and Constructive Group Thinking Sustainability (CGTS) (Hu, 2021).

RGTS refers to the degree of reciprocity of a talk sequence. Stahl (2014) argued that effective group cognition requires long sequences of responses involving a group of individuals who jointly construct a series of adjacent pairs to achieve a larger cognitive goal. Similarly, GTS assumes that a reciprocal sequence of conversation is fundamental to the development of sustained group thinking. PGTS and CGTS are measures of the efficacy of reciprocal group thinking, where PGTS represents the extent to which a group engages in high-order thinking and CGTS represents the extent to which a group constructs new knowledge.

In GTS, the thinking trajectory in a talk sequence is coded as a three-dimensional binary vector of reciprocity (R), productivity (P), and constructiveness (C): $\{(R_1, P_1, C_1), (R_2, P_2, C_2), (R_3, P_3, C_3), \ldots, (R_n, P_n, C_n)\}$. The average number of turns in a reciprocal turn-taking sequence, productive talk moves in reciprocal talk sequences, and new ideas in reciprocal talk sequences are operationalized as RGTS, PGTS, and CGTS, respectively (Hu, 2021). It is important to note that qualitative analysis of group discussions is also recommended to complement the quantitative analysis of the GTS (Hu, 2021; Wegerif et al., 2017). GTS enables analysis of the performance of a single group across a wide range of tasks and comparison of the performance of multiple groups on the same task. As GTS could examine group interaction patterns in a group thinking test, GTS results are an excellent complement to a GTM score.

3.4.3. Mathematics self-concept

The measurement of Mathematics self-concept was accomplished through a four-point Likert scale consisting of nine items adapted from the TIMSS 2015 questionnaire for fourth graders in Taiwan (Hooper et al., 2013). The participants indicated their level of agreement with each statement, with the scale ranging from 1 (*strongly agree*) to 4 (*strongly disagree*). amongst the nine items, four are positive and five are reverse scored. This scale showed high-level internal consistency (Cronbach's alpha = 0.86, $n = 277$). The exploratory factor analysis indicated a 2-factor structure, rather than the original one-factor model. The confirmatory factor analysis further demonstrated a good fit of the 2-factor model ($\chi^2(26) = 62.8$, $p < .001$, CFI = 0.96, TLI = 0.95, RMSEA=0.072, 90%CI [0.049, 0.094], SRMR = 0.050) for the adapted scale.

3.4.4. Mathematics enjoyment

Mathematics learning enjoyment was assessed using a nine-item, four-point Likert scale adapted from the TIMSS 2015 questionnaire for fourth graders in Taiwan (Hooper et al., 2013), similar to the mathematics self-concept scale. Two cases were excluded due to the same responses to all nine items. The scale achieved a high degree of internal consistency, with Cronbach's alpha achieving 0.87 ($n = 279$). CFA of the original one-factor model indicated a good fit for the data ($\chi^2(27) = 51.06$, $p < .01$, CFI = 0.98, TLI = 0.97, RMSEA = 0.057, 90%CI [0.032 0.080], SRMR = 0.033).

3.4.5. Social anxiety

Social anxiety was measured using the 10-item Chinese version of the Social Anxiety Scale for Children–Revised (La Greca & Stone, 1993), adapted to a four-point Likert scale (1 = *strongly agree*, 2 = *somewhat agree*, 3 = *somewhat disagree*, and 4 = *strongly disagree*). The scale showed a high level of internal consistency in this study, with Cronbach's alpha achieving 0.86 ($n = 281$). CFA also indicated a good fit of the original one-factor 10-item model for the data ($\chi^2(35) = 62.12$, $p < .01$, CFI = 0.97, TLI = 0.96, RMSEA = 0.053, 90%CI [0.030 0.073], SRMR = 0.044).

3.4.6. Perspective taking

Perspective-taking ability was measured using the subscale on perspective-taking in the empathy measure developed by Davis (1983), which contained seven items, two of which were reversed. The items were translated into Chinese and required students to report on a four-point Likert scale (1 = *strongly agree*, 2 = *somewhat agree*, 3 = *somewhat disagree*, and 4 = *strongly disagree*). EFA suggested a 5-item, one-factor model. The internal consistency of Cronbach's alpha for the adapted scale was 0.65, which was slightly below the set cutoff of 0.70. Given the limited number of items in this scale, the slightly low internal consistency was deemed acceptable in this study. CFA further indicated an acceptable model fit of this revised one-factor model ($\chi^2(5) = 10.29$, $p = .068$, CFI = 0.96, TLI=0.92, RMSEA=0.087, 90%CI [0.000 0.162], SRMR = 0.042).

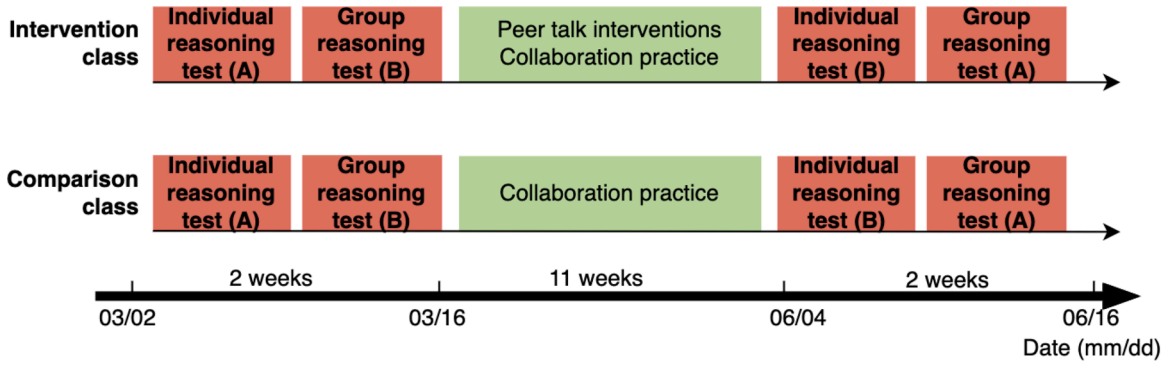


Fig. 2. Study procedures for two classes.

3.5. Procedure

Oral consent was obtained from the students and written consent forms were collected from the principal, the two teachers, and the students' guardians, before the start of the course. Fig. 2 shows a comparison between the study procedures for the two classes. Both classes of students took pre-test and post-test individual and group thinking tests; this was achieved by switching the two sets of puzzles. There was a 1-week interval between the individual and group tests.

At the beginning of the study, we collected the students' recent mathematics and Chinese grades from their teachers. The students were also required to name three classmates with whom they would like to be grouped and three classmates they most frequently referred to for help with mathematics problems. To optimize the comparability of the groups, the students in each class were categorized into three levels based on their total scores for mathematics, Chinese, and individual thinking skill. One low level, one mid-level, and one high-level student was assigned to each group, with the mid-level students sitting in the middle of each group. To motivate low-performing students to participate, we prioritized their grouping preferences when choosing the mid-level and high-level members of their groups. We attempted to ensure that there was one girl and one boy in each group; however, owing to the imbalanced sample, there were three all-boy groups in the intervention class. The grouping results were adjusted slightly based on suggestions from teachers concerning the styles and overall mathematical ability of various groups. We also made small adjustments to the grouping in response to strong personal feedback from some students (e.g., that there were very poor personal relationships between the members of a group). After the grouping, the students in both classes completed some instruments relevant to group thinking that measured their mathematics self-concept, mathematics learning enjoyment, social anxiety, and perspective-taking ability.

During the study, the two classes were designed with different pedagogical approaches. The first author took charge of the design and implementation of both versions of the course. Although she had limited part-time tutor experience in after-school tutoring agencies, she lacked full-time teaching experience in primary school. Nevertheless, she was well-versed in the talk tools and design of iTalk-iSee used in the study. The second author acted as the teaching assistant responsible for overseeing the data collection system and usage of iTalk-iSee, and also assisted with pedagogical design and classroom management. He possessed over a decade of experience in software development and six years of full-time teaching experience at a junior high school.

In the intervention class, students participated in intervention sessions interspersed with practice sessions, meeting with us (i.e., the first two authors) twice a week throughout a semester. A typical practice session consisted of two major activities: reviewing previously learned content and engaging in a new task for around 15 to 20 min. We guided students to recall and share their understanding of talk virtues or tools learned in previous sessions and introduced additional examples to clarify any misunderstandings or demonstrate the proper use of the talk tools. In a typical intervention session, there were four activities. The first involved evaluating the problem-solving performance of the previous task, discussing and reflecting on solutions, and demonstrating typical problem-solving challenges and strategies. The second evaluated the talk performance of the previous task, reflecting on issues that arose and discussing how to address them. The third involved learning new knowledge on productive peer talk, with a focus on a specific talk issue illustrated by a pre-recorded instructional video, followed by reflection and discussion amongst students. The final activity involved the hands-on use of iTalk–iSee to analyse and reflect on the talk performance about the newly learned topic from the previous task. For example, when students used iTalk–iSee to analyse their usage of I-Talk tools, they collaboratively coded I-Talk tools in each utterance, discussed visualizations that illustrated their performance against the teacher-set standard, analysed the usage rates of each talk tool and the individual usage of talk tools across the problem-solving process, and reflected on their overall performance and discussed possible plans to improve their use of I-Talk tools.

In contrast, we met with students in the comparison class only once a week and provided them only with practice sessions. These sessions followed a similar structure to those in the intervention class, with evaluations of problem-solving and group talk performance and collaborative engagement in new tasks. However, students in the comparison class did not learn how to talk with peers or have access to the talk analysis tool iTalk-iSee until the final make-up sessions, which was done in accordance with ethical considerations.

Table 3
Coding protocol for group thinking sustainability.

Dimension	"1″	"0″	"NULL"
Reciprocal	The turn contains feedback to the previous speaker, including simple feedback (such as brief acknowledgement and repetition) and productive feedback (such as elaboration, justification, evaluation, and challenge).	The turn does not contain feedback to the previous speaker and initiates a new thinking trajectory.	The turn contains no information that is relevant to the current problem-solving activity. The regulation of time allocation or off-task behaviour is on-task.
Productive	The turn contains at least one productive talk move.	The turn does not contain any productive talk moves.	
Constructive	The turn contributes at least one type of new content knowledge to the current task, such as new ideas, new inferences, new speculations, or new proposals.	The turn does not contain relevant content knowledge, or it refers to or repeats previously contributed content knowledge.	

3.6. Data sources

3.6.1. Changes of thinking test scores

To examine whether the intervention class exhibited better group thinking than the comparison class, we used the GTM to quantitatively test group thinking at pre and post-test. We then used SPSS software version 26 (IBM Corp., Armonk, NY, USA) to quantitatively analyse the group and individual thinking scores in the two classes. We compared the pre–post difference between the two classes using an independent samples *t*-test rather than an analysis of covariance (ANCOVA) because the tests were identical at pre- and post-test and the initial group thinking score affected the grouping of students (Wright, 2006).

3.6.2. Changes of collaborative discourse

To complement the understanding of test scores, we also examined the quality of collaborative discourse in doing the tests from the aspects of interaction intensity, participation inequality, the usage rate of talk tools, and GTS to complement the thinking test scores. To determine interaction intensity, we examined the total number of turns a group performed in solving the puzzles. The participation inequality of a group was calculated as the standard deviation of its members' individual participation rates.

The usage rates of four types of peer talk tools (I-Talk, You-Talk, We-Talk-Equity, and We-Talk-Convergence) in the two classes were calculated as the accumulated usage rates of each type's individual talk tools. For example, the usage rate of I-Talk was the sum of the usage rates of the "share information," "express new idea," "explain oneself," "build on oneself," and "self-reflect" tools.

Four undergraduate volunteers were recruited for the coding project and received satisfactory compensation for their work. They were not informed about the intervention but were provided with a detailed introduction to all codes. All potential coders underwent sufficient training before they became qualified for formal coding tasks. The training process lasted for approximately two weeks, while the formal coding process lasted for about three months (refer to Fig. A2 in the Appendix). Three qualified coders were involved in labelling the 18 talk tools turn by turn.

Fuzzy kappa— a modified version of Cohen's kappa that allows each unit of analysis to be assigned multiple codes (Kirilenko & Stepchenkova, 2016)—was used to measure the inter-coder agreement on the coding of talk tools. The indicators for any pair of coders were all satisfactory (fuzzy kappa values > 0.60).

To determine GTS, the coding for the productivity of each turn was labelled based on whether the turn involved the use of any talk tool (see Table 3). A turn was also coded as reciprocal if it contained any form of feedback to the previous speaker and as constructive if it contributed at least one type of new content knowledge to the discussion (Hu, 2021). The three trained coders first practiced coding four identical group discussions that each contained approximately 200 turns in each. The average Cohen's kappa results for reciprocity and constructiveness amongst any pair of coders were adequate (both > 0.80). The three coders then separated the coding workload and thereafter independently finished coding all of the groups.

3.6.3. Processes and dynamics of changes

To understand how the teaching-talk-for-thinking program may influence group thinking, we conducted a case study to illustrate the processes and dynamics of change in group thinking. Following the logic of enquiry of interactional ethnography (Bridges et al., 2020), we compared how value-added triadic groups in the intervention class and the comparison class improved their grouping values from pre- to post-test. Interactional ethnography helps researchers to identify rich points (Agar, 2006) of interest and confusion that can serve as anchors for decomposing the complexity of observed phenomena. An interactional ethnography selects telling cases to reveal non-linear interaction dynamics; it also uses event mapping to construct a graphical representation of events to enable the focused analysis of rich points (Bridges et al., 2020; Green & Bridges, 2018). An event map is a visual aid that assists in creating a historical framework for specific events. It displays events and sub-events and allows for the zooming in on details to facilitate comparisons or patterns.

After thoroughly examining the collaborative efforts of all groups, we selected Group 8 (G8) from the intervention class and Group 38 (G38) from the comparison class as telling cases. These groups were chosen because they were able to demonstrate the differences between the two classes in terms of their improvement in grouping values. The three members of G8 were Qi, Liu, and Li. Qi was the

Table 4
Background information of members of G8.

	Li	Liu	Qi
Age	9	10	10
Gender	F	M	F
Recent mathematics grade[a]	60	97	110
Recent Chinese grade[a]	76	103	116
Mathematics self-concept[b]	1.56	3.56	3
Mathematics enjoyment[b]	3.67	3.89	3.78
Social anxiety[b]	1.3	1.6	2.1
Perspective-taking[b]	3.4	3.6	3.2
Family resources[c]	3	15	3

Note.
[a] The maximum score was 120.
[b] Measured on a 4-point Likert scale with a maximum score of 4.
[c] One point was given for each of 19 categories of family resources (e.g., individual desk, individual room, individual computer/pad, guest bedroom, books, and musical instruments), for a maximum score of 19.

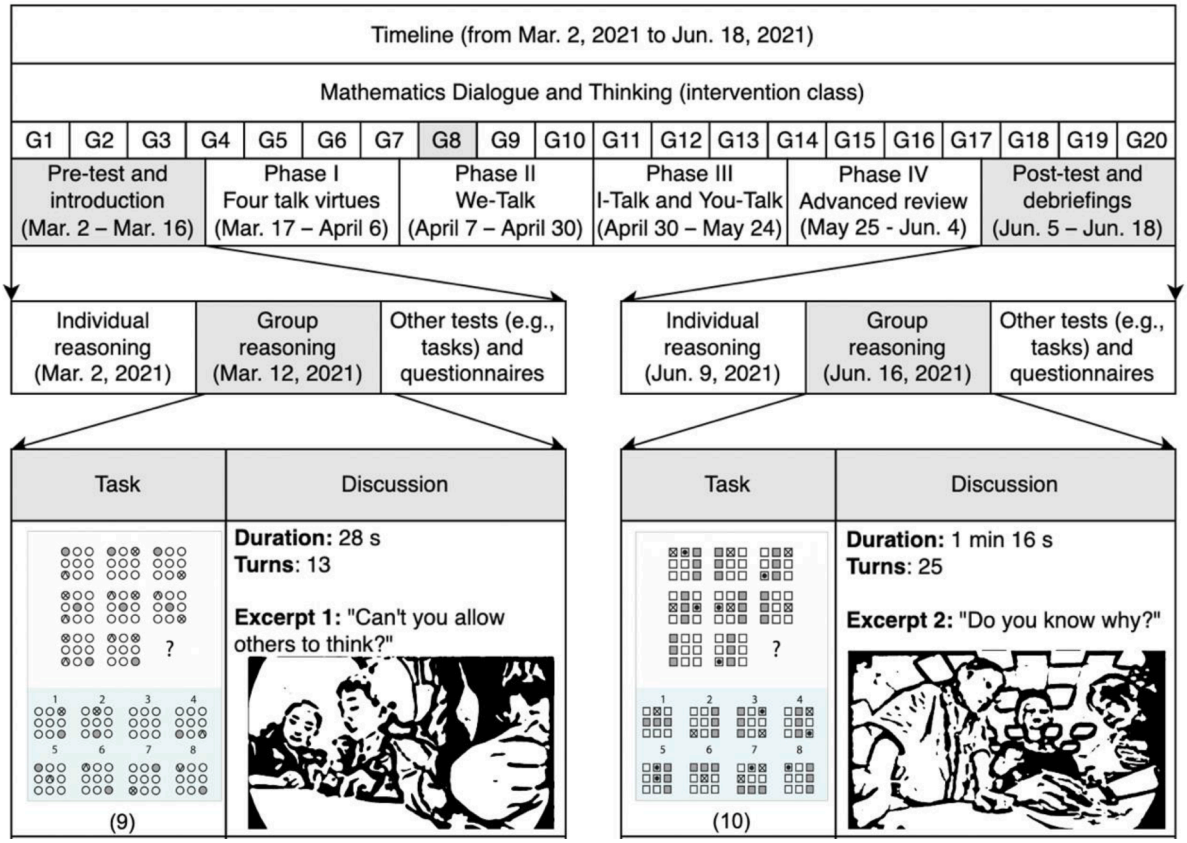


Fig. 3. Event map: puzzle-solving of G8 at pre- and post-test.

most academically gifted but also the most socially anxious (see Table 4). Liu achieved the second-highest mathematics scores but he enjoyed learning mathematics the most and had the highest self-concept in mathematics. Liu also had the highest perspective-taking ability and the most family resources. Li was the least academically gifted in G8 and had the lowest mathematics self-concept. However, she greatly enjoyed learning mathematics and was the least socially anxious member of the group. Furthermore, before the intervention, Li had reported that she wanted to be in a group with Qi.

To illustrate how group thinking changed in G8, we contrasted their collaborations on two similar puzzles that they failed to solve one at pre-test but solved the other at post-test. These two puzzles were equivalent in patterns and have a comparable level of difficulty. The event map in Fig. 3 anchors the puzzles given to G8 in the intervention program.

The members of G38 were Jiang, Tang, and Zhou. Jiang was the most academically gifted but also the most socially anxious (see Table 5). Although he enjoyed learning mathematics, he had the lowest self-concept in mathematics in the group and also the lowest

Table 5
Background information of members of Group 38.

	Jiang	Tang	Zhou
Age	9	10	10
Gender	M	F	F
Recent mathematics grade[a]	102	66	87
Recent Chinese grade[a]	87	86	100
Mathematics self-concept[b]	2.56	2.67	3.67
Mathematics enjoyment[b]	3.67	3.56	3.56
Social anxiety[b]	2.3	1.5	1.8
Perspective-taking[b]	2.6	3.4	3
Family resources[c]	1	1	18

Note.
[a] The maximum score was 120.
[b] Measured on a 4-point Likert scale with a maximum score of 4.
[c] One point was given for each of 19 categories of family resources (e.g., individual desk, individual room, individual computer/pad, guest bedroom, books, and musical instruments), for a maximum score of 19.

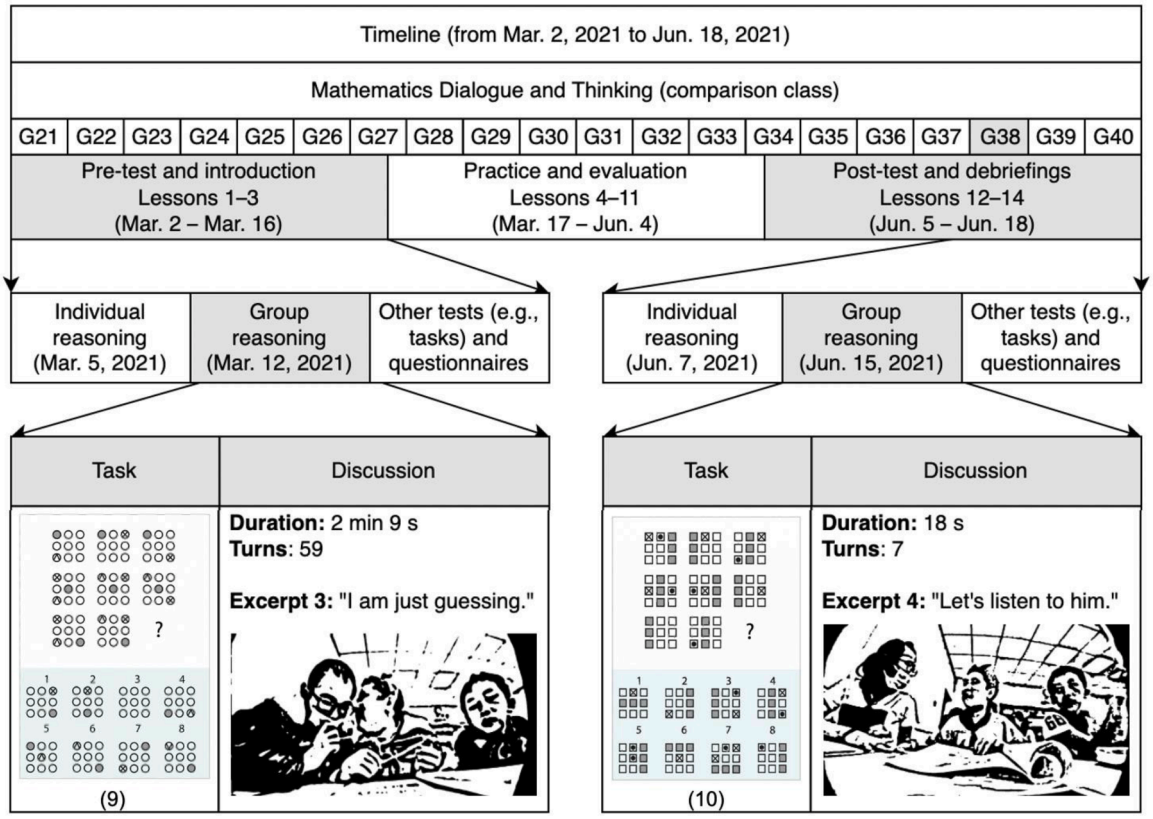


Fig. 4. Event map: puzzle-solving of G38 at pre- and post-test.

perspective-taking ability of the three members. Tang was the least academically gifted; however, she enjoyed learning mathematics and had the lowest social anxiety and highest perspective-taking ability in the group. Zhou also enjoyed learning mathematics and had the highest self-concept in mathematics and the highest number of family resources. Furthermore, before the intervention, Zhou had reported that she wanted to be in a group with Tang.

To illustrate how group thinking changed in G38 and compare G38 to G8, we contrasted the collaboration of G38 on the same pair of puzzles. The event map in Fig. 4 situates these pre- and post- puzzles given to G38 in the overall program.

Table 6
Descriptive statistics for grouping value in the two classes.

	Class	*n*	*M*	*SD*	*Min*	*Max*
Pre-test	Intervention	20	−0.30	1.92	−3.00	3.00
	Comparison	20	0.40	2.09	−3.00	4.00
Post-test	Intervention	20	−0.25	2.00	−4.00	3.00
	Comparison	20	0.15	2.11	−6.00	4.00
Pre-post difference	Intervention	20	0.05	2.42	−4.00	4.00
	Comparison	20	−0.25	2.69	−5.00	6.00

Table 7
Number of groups at various levels of grouping value in the two classes.

Test	Class	Value-added	Value-detracted	Value-neutral
Pre-test	Intervention	4	6	10
	Comparison	3	3	14
Post-test	Intervention	5	3	12
	Comparison	4	3	13

Table 8
Changes in the level of grouping value in the two classes from pre-test to post-test.

Class	Improved	Maintained	Decreased
Intervention	7	9	4
Comparison	4	12	4

Table 9
Number of turns exhibited by the two classes in pre- and post- group thinking tests.

Class	Test	*n*	*M*	*SD*	*Min*	*Max*
Intervention	Pre	20	220	98	92	369
	Post	20	112	50	30	213
	Paired *t*-test	$t(19) = 5.587, p < .001$				
Comparison	Pre	20	238	73	74	337
	Post	20	91	41	15	164
	Paired *t*-test	$t(19) = 8.707, p < .001$				
Between-group difference		$F(1, 37) = 3.349, p = .075$				

4. Results

4.1. Was group thinking improved in the intervention class relative to the comparison class? (RQ1)

4.1.1. Changes in the group thinking measure (GTM) scores

The comparison class had a larger average grouping value (i.e., the difference between the group score and the highest score of an individual member) than the intervention class at both pre- and post-test (see Table 6). On average, there was an increased grouping value from pre-test to post-test in the intervention class and a reduced grouping value from pre-test to post-test in the comparison class. However, an independent samples *t*-test revealed that there were no significant differences between the two classes in terms of pre-test grouping value, post-test grouping value, or pre–post difference in grouping value (all $ps > 0.05$). A paired samples *t*-test indicated that there were no significant differences between the pre- and post-test grouping value in either class (all $ps > 0.05$).

In terms of the levels of grouping value, more than half of the groups were value-neutral groups (see Table 7). The intervention class had one more value-added group and three fewer value-detracted groups at post-test than at pre-test, whereas the comparison class only had one more value-added group at post-test than at pre-test.

We also sorted groups into three categories based on whether they improved, maintained, or decreased their level of grouping value from pre-test to post-test. Seven groups in the intervention class and four groups in the comparison class improved by at least one level of grouping value, whereas four groups in each class decreased their level of grouping value (see Table 8).

4.1.2. Discourse of group thinking test

In this section, we report the results of a comparison between the two classes' collaborative discourse in performing group thinking tests in terms of interaction intensity, participation inequality, the usage rate of talk tools, and GTS.

4.1.2.1. Interaction intensity and participation inequality. There were 15 visual puzzles in each thinking test. Across the two classes, the students exhibited an average of 229 ($SD = 86$) and 101 ($SD = 47$) turns at pre- and post-tests, respectively. There were no significant

Table 10
Participation inequality for the two classes in pre- and post- group thinking tests.

Class	Test	*n*	*M*	*SD*	*Min*	*Max*
Intervention	Pre	20	0.126	0.102	0.02	0.29
	Post	20	0.101	0.076	0.03	0.33
	Paired *t*-test	$t(19) = 1.074, p = .296$				
Comparison	Pre	20	0.112	0.073	0.01	0.29
	Post	20	0.117	0.070	0.04	0.24
	Paired *t*-test	$t(19) = -0.243, p = .811$				
Between-group difference		$F(1, 37) = 0.656, p = .423$				

Table 11
Descriptive statistics of the usage rates of talk tools.

	Class	*n*	I-Talk (%)	You-Talk (%)	We-Talk-Equity (%)	We-Talk-Convergence (%)
Pre-test	Intervention	20	32.3 (6.4)	22.7 (8.6)	5.9 (3.2)	10.2 (4.4)
M (SD)	Comparison	20	34.0 (7.2)	24.0 (5.5)	5.0 (3.0)	9.3 (3.5)
Post-test	Intervention	20	36.3 (8.6)	31.9 (13.6)	9.9 (6.8)	6.1 (3.4)
M (SD)	Comparison	20	38.9 (10.8)	25.2 (9.7)	5.0 (3.1)	7.9 (5.1)

Table 12
Pre- and post-test group thinking sustainability of the two classes.

Test	Class	*n*	RGTS/*M(SD)*	PGTS/*M(SD)*	CGTS/*M(SD)*
Pre	Intervention	20	2.59 (0.73)	1.71 (0.74)	0.88 (0.57)
	Comparison	20	2.24 (0.34)	1.47 (0.30)	0.72 (0.20)
	t-test		$t(38) = 1.921, p > .05$	$t(38) = 1.312, p > .05$	$t(38) = 1.189, p > .05$
Post	Intervention	20	2.55 (0.69)	1.97 (0.80)	0.99 (0.45)
	Comparison	20	2.15 (0.37)	1.46 (0.30)	0.82 (0.24)
	t-test		$t(38) = 2.264, p < .05$	$t(38) = 2.711, p < .05$	$t(38) = 1.556, p > .05$

Note. RGTS = reciprocal group thinking sustainability, PTGS = productive group thinking sustainability, CGTS = constructive group thinking sustainability.

differences between the two classes in the total number of turns in either the pre-test or post-test thinking tasks (see Table 9). The results of a one-way ANCOVA indicated that there was no significant difference between the two classes in the number of turns at post-test after adjusting for the number of turns at pre-test. Paired samples *t*-tests revealed that both the intervention class and the comparison class produced significantly less turns at post-test than at pre-test.

Regarding participation inequality, there were no significant differences between the two classes at post-test after adjusting for the pre-test difference (see Table 10). Paired samples *t*-tests also indicated no significant changes from pre-test to post-test for either class (all $ps < 0.05$).

4.1.2.2. Usage rates of talk tools. The descriptive statistics on the usage rates of the four types of talk tools are presented in Table 11. Shapiro–Wilk's tests indicated that all of the usage rates of these four types of talk tools for the two classes in pre- and post-test thinking tasks were approximately normally distributed ($p > .05$), except for the pre-test usage rate of We-Talk-Equity by the comparison class and the post-test usage rate of I-Talk by the intervention class ($ps < 0.05$). Therefore, we assumed that all of the variables on the usage rates of the four talk tools were part of a normal distribution and adopted parametric analyses for examining the between- and within-group differences.

One-way ANCOVAs indicated that the post-test usage rate of I-Talk was not significantly different between the two classes after adjusting for the pre-test rate. The difference in post-test usage rate of You-Talk between the two classes was also non-significant but with a *p*-value close to 0.05 after adjusting for the pre-test difference, $F(1, 37) = 3.461$, $p = .071$, partial $\eta^2 = 0.086$.

There was not a significant difference between the post-test usage rate of We-Talk-Convergence of the two classes after adjusting for the pre-test usage rate, but there was a significant difference between the post-test usage rate of We-Talk-Equity of the two classes after adjusting for the pre-test usage rate, $F(1, 37) = 8.665$, $p < .01$, partial $\eta^2 = 0.190$.

The results of paired samples *t*-tests to examine within-group differences in the usage rates of the talk tools indicated that the intervention class made significantly more use of You-Talk ($t(19) = -2.851, p < .05$) and We-Talk-Equity ($t(19) = -2.251, p < .05$) but significantly less use of We-Talk-Convergence ($t(19) = 3.521, p < .01$). There was no significant difference in the use of I-Talk by the intervention class between the pre- and post-test thinking tasks.

The paired samples *t*-test also indicated that the comparison class made significantly more use of I-Talk ($t(19) = -2.567, p < .05$) at post-test than at pre-test. There was no significant difference in the usage rates of other talk tools amongst the comparison class between the pre- and post-test group thinking tasks (all $ps > 0.05$).

Table 13
Background information of two telling cases in the two classes.

	G8		G38	
	Pre-test	Post-test	Pre-test	Post-test
Group thinking	8	13	8	12
Individual thinking	Qi/Liu/Li	Qi/Liu/Li	Jiang/Tang/Zhou	Jiang/Tang/Zhou
	7/8/7	9/10/7	8/9/11	8/9/8
Turns	215	213	333	86
Participation inequality	0.1	0.08	0.13	0.18
RGTS	1.95	2.44	2.04	2.67
PGTS	1.43	1.71	1.03	1.93
CGTS	0.9	1.15	0.53	0.8
I-Talk	0.4	0.37	0.29	0.31
You-Talk	0.24	0.33	0.15	0.19
We-Talk-Convergence	0.08	0.05	0.07	0.15
We-Talk-Equity	0.05	0.1	0.02	0.05

Note. RGTS = reciprocal group thinking sustainability, PGTS = productive group thinking sustainability, CGTS = constructive group thinking sustainability.

4.1.2.3. Group thinking sustainability (GTS). The intervention class performed slightly better on average than the comparison class in the three-level metrics of GTS in both the pre- and post-tests (see Table 12). Independent samples *t*-tests indicated that there were no significant differences in pre-test GTS between the two classes, but the intervention class had significantly higher RGTS and PGTS than the comparison class at post-test. That is, the intervention class showed longer sequence of reciprocal dialogue than the comparison class and the reciprocal dialogue sequence involves more productive peer talk moves.

One-way ANCOVAs indicated that there was a significant difference between the two classes in post-test RGTS after adjusting for pre-test RGTS, $F(1, 37) = 5.064$, $p < .05$, partial $\eta^2 = 0.120$. There was also a significant difference between the two classes in post-test PGTS after adjusting for pre-test PGTS, $F(1, 37) = 6.928$, $p < .05$, partial $\eta^2 = 0.158$. However, there was no significant difference between the two classes in post-intervention CGTS after adjusting for pre-intervention CGTS.

4.2. How did the intervention help to improve group thinking? (RQ2)

As shown in Table 8, seven groups in the intervention class and four groups in the comparison class increased their level of grouping value after the intervention. This section examines whether the improvement of these groups was related to the intervention and, if so, how the intervention affected the grouping value differently in the two classes.

G8 in the intervention class and G38 in the comparison class were chosen as telling cases to illustrate the differences between the two classes in terms of how groups improved their grouping values. These two groups achieved similar scores in pre- and post-test individual and group thinking tasks (see Table 13). The grouping value for G8 improved from 0 at pre-test to 3 at post-test and for G38 improved from −3 at pre-test to 3 at post-test. Both groups improved their GTS and increased their usage rates of talk tools from pre-test to post-test. The participation of G8 became more equal at post-test whereas that of G38 became more unequal at post-test. G38 also exhibited a drastic decrease in the total number of turns from pre-test to post-test (from 333 to 86).

In the following sections, we closely examine the puzzle-solving processes of these two groups and compare their performances on two equivalent puzzles at the pre- and post-test respectively. This reveals how the groups improved their grouping value in the study and whether the mechanism of improvement differed between the two classes.

4.2.1. G8 in the intervention class

4.2.1.1. Comparison of performance on the selected pair of puzzles. When solving Puzzle 9 at pre-test, Liu and Qi asked Li to express her ideas because Li complained that she had not been given an opportunity to speak (#79, #80 in Excerpt 1). However, Li offered no ideas about how to solve the puzzle (#82). Liu then began to share his idea, which caused Li to interrupt and complain "can't you allow others to think?" (#84). However, her complaint was ignored. Qi also began to share her idea (#85), while Liu continued sharing his idea. Subsequently, Li seemed to think of an idea and interrupted Qi and Liu to explain it, but she again failed to express her thoughts clearly (#88). Liu then wrote down his answer without further explanation, even though Qi was still voicing her idea (#87). His-last comment ("What a slow reaction!"; #90) was a negative observation about his team members' performance and revealed his authoritative role in the group. Overall, this excerpt indicates that there was some tension in the relationships between the members of G8, such that they focused on their own thinking. Li was ignored by the other members and Liu was too impatient to wait for Li or listen to Qi or to give sufficient explanations.

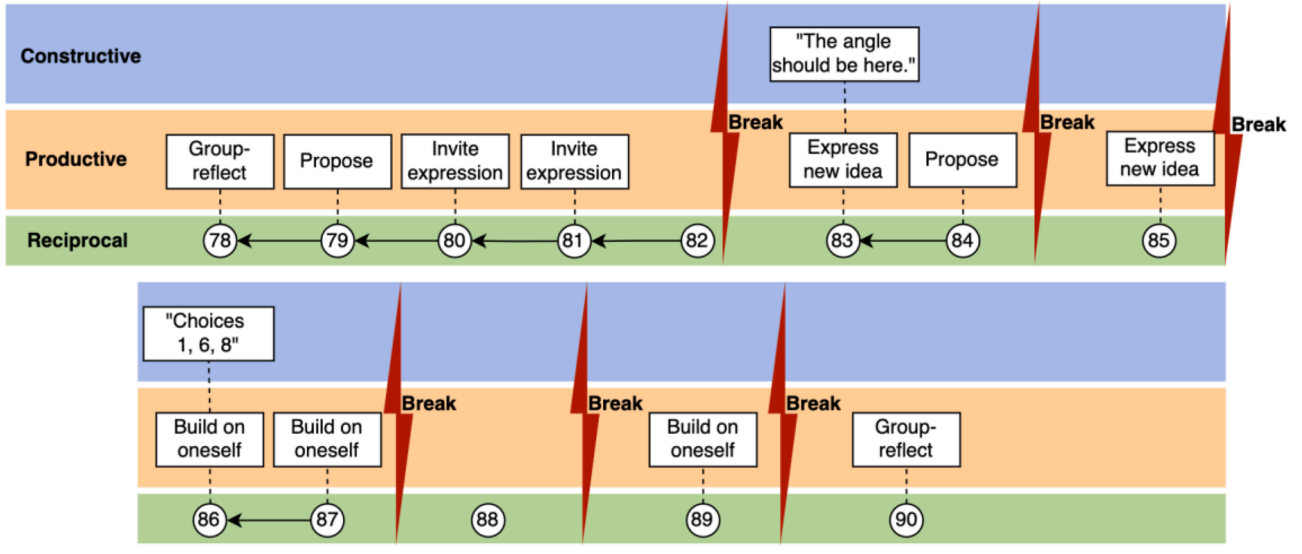


Fig. 5. Dynamics of GTS in G8 during the pre-test Puzzle 9. The numbers in the circles represent the turn order and the arrows show the response to the previous turn. A lightning symbol represents a break in the group thinking flow, meaning there was no response. The rectangles show the use of talk tools and the dotted lines connect new ideas and talk tools to the turns they belong to.

Excerpt 1 "Can't you allow others to think?"

Turn	Start time	End time	Speaker	Content	Embodied actions
78	0:04:36	0:04:38	Li	I knew the answer. You did not allow me to speak. You just argued with each other.	
79	00:04:38	00:04:39	Qi	Then let her speak this time. Let her speak this time.	#79. Qi (left) looked at Liu (middle) and pointed at Li (right) while speaking. Li looked at Qi. Liu smiled and looked at the paper.
80	00:04:39	00:04:40	Liu	Come! you speak.	
81	00:04:40	00:04:40	Qi	You speak!	
82	00:04:41	00:04:46	Li	I finish speaking… (muffled, not clear). Finished…Oh… right.	
83	00:04:49	00:04:53	Liu	Ah! Isn't this simple! Whatever, the angle should be here. Exclude… Only choice 1!	
84	00:04:53	00:04:54	Li	Can't you allow others to think?	#85. Qi (left) pointed at the paper while speaking. Liu (middle) looked at the paper and spoke at the same time. Li (right) looked at the paper.
85	00:04:54	00:04:56	Qi	This is here. This is here.	
86	00:04:55	00:04:56	Liu	Choices 1, 6, 8… Only 1, 6, 8.	
87	00:04:57	00:04:58	Qi	Then this is this corner.	
88	00:04:56	00:05:00	Li	Ah! I know! I know! Excuse me… (muffled, not clear).	
89	00:05:01	00:05:02	Qi	Then this is… this cross.	#88. Li (right) laughed and used her hand to move away Qi's (left) hand on the paper while speaking. Both Qi and Liu (middle) looked at the paper.
90	00:05:02	00:05:04	Liu	What a slow reaction!	

The thinking trajectory in Excerpt 1 could be expressed as {… (0, 1, 0), (1, 1, 0), (1,1, 0), (1, 1, 0), (1, 0, 0), (0, 1, 1), (1, 1, 0), (0, 1, 0), (0, 1, 1), (0, 1, 0), (0, 0, 0), (0, 1, 0), (0, 1, 0)…}. It was divided into seven subsequences (see Fig. 5).

The GTS scores in Excerpt 1 were RGTS = 13/7 = 1.86, PGTS = 11/7 = 1.57, and CGTS = 2/7 = 0.29. There were fewer than two consecutive reciprocal turns on average, which indicates rather fragmented group thinking. Although the students demonstrated a reasonably high usage rate of peer talk tools, they produced very few new ideas. In particular, they made most use of the I-Talk tools, including "express new idea" and "build on oneself," indicating the cognitive dominance of a single member and a low level of intergroup thinking.

When solving an equivalent puzzle (Puzzle 10) at post-test, Liu initially blamed Li for voicing an answer without providing explanations (#96). This indicated an increased awareness in G8 in their post-test discussion of the value of persuading others by explaining or justifying one's ideas rather than just pursuing an answer. Compared with the pre-test, at post-test Liu was more willing

to ask whether the other members understood his idea and to provide explanations (#105, #115). Although he sometimes blamed Li (#96) or ignored her voice (#106), Liu was relatively more responsive to Li (#99, #101) than at pre-test. There was also good constructive discussion between Liu and Qi at post-test (#102–#103, #107–#119).

Excerpt 2. "Do you know why?"

Turn	Start time	End time	Speaker	Content	Embodied actions
96	0:06:18	0:06:25	Liu	She just stated an answer every time. Come on! You! Look at this problem.	
97	0:06:26	0:06:32	Li	Um… this is… this is… eek… this should be three.	#97. Li (right) held up three fingers and looked at the paper while speaking. Both Qi (left) and Liu (middle) looked at the paper.
98	0:06:33	0:06:34	Qi	This…	
99	0:06:35	0:06:35	Liu	Wrong!	
100	0:06:35	0:06:35	Li	It is three.	
101	0:06:35	0:06:36	Liu	I told you it was wrong.	
102	0:06:39	0:06:42	Qi	I think we should look at this vertically.	
103	0:06:42	0:06:43	Liu	We should look at this vertically.	
104	0:06:44	0:06:44	Li	Which one do you think?	
105	0:06:45	0:06:55	Liu	Come! First, we all know the route of the black. Do you know why? I tell you why. The black moves along the diagonal line.	
106	0:06:55	0:06:56	Li	Ah! Wait! This should be three!	#105. Liu (middle) used a pen to point at figures on the paper while speaking. Both Qi (left) and Li (right) looked where he pointed.
107	0:06:55	0:07:07	Liu	When it comes to this edge, it begins here, then here. First, we know… it comes here. Should it come back here. So, 8, 8, 2, 5.	
108	0:07:08	0:07:08	Qi	8, 2, 5 are okay.	
109	0:07:08	0:07:09	Liu	8, 2, 5are okay.	#112. Qi (left) pointed at the paper while speaking. Both Liu (middle) and Li (right) looked where she pointed.
110	0:07:09	0:07:10	Qi	Then look at the others.	
111	0:07:11	0:07:12	Liu	Then let's look at the others.	
112	0:07:12	0:07:14	Qi	It can't be two circles in the middle.	
113	0:07:15	0:07:20	Liu	Exclude 5. This square… you look at this square. Does it move down one grid every time?	
114	0:07:21	0:07:21	Qi	Yes.	
115	0:07:23	0:07:25	Liu	So, so it can't be this. Do you know why?	
116	0:07:27	0:07:28	Qi	It should be 8.	
117	0:07:28	0:07:28	Liu	Yes.	
118	0:07:29	0:07:31	Qi	8 covers the square.	
119	0:07:32	0:07:33	Liu	Yes! It covers the square.	
120	0:07:34	0:07:34	Li	Oh, I see.	

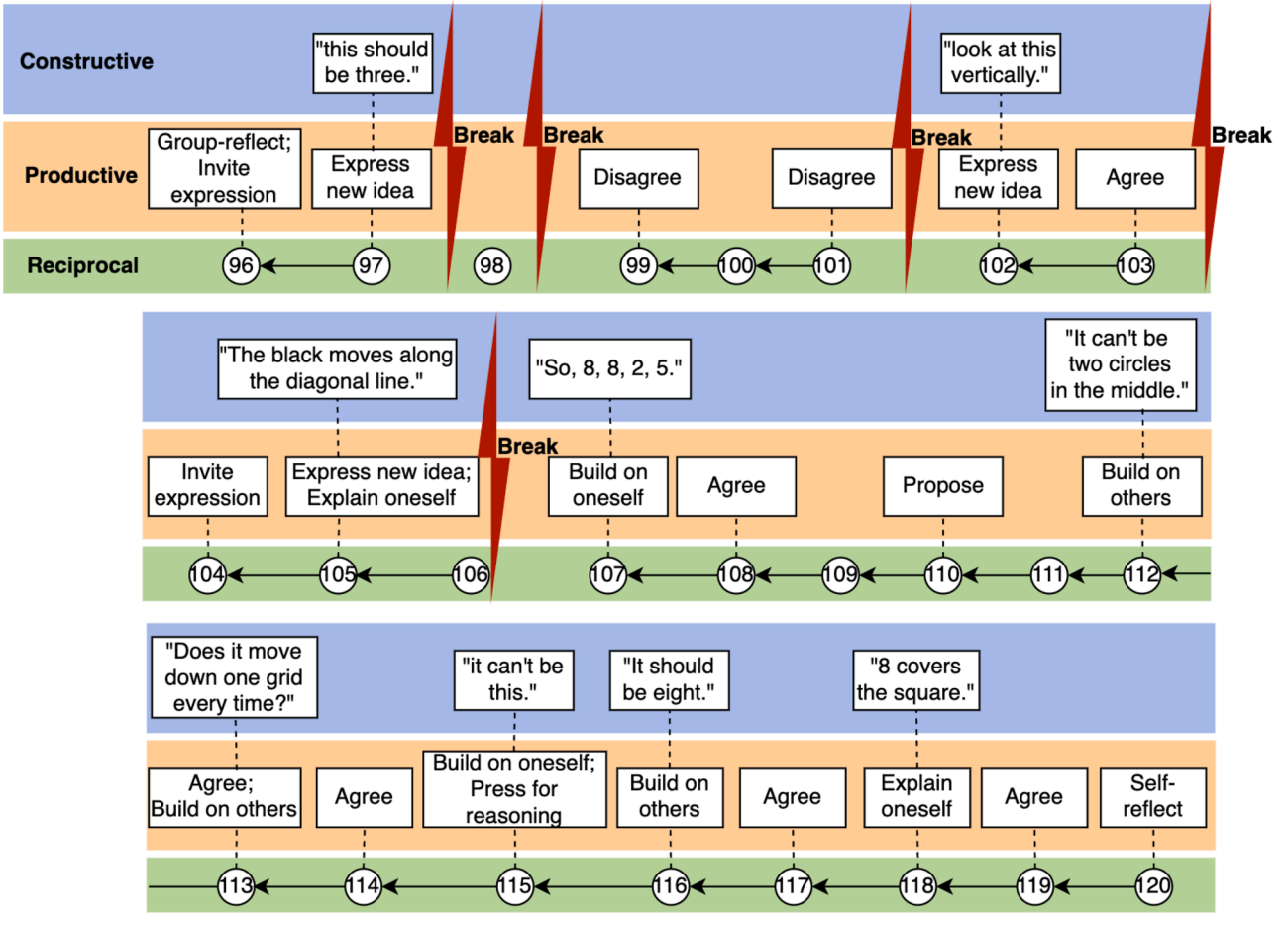


Fig. 6. Dynamics of GTS in G8 during the post-test Puzzle 10.

Excerpt 3. "I'm just guessing."

Turn	Start time	End time	Speaker	Content	Embodied actions
231	00:11:49	00:11:52	Tang	Change… The black dots change.	 #240. Tang (middle) whispered in Zhou's (left) ear. Both laughed. Jiang (right) looked at the paper.
232	00:11:53	00:11:53	Jiang	Correct.	
233	00:11:54	00:12:01	Zhou	Correct. So, the last one should be this. You look. Then it is… Then like this…	
234	00:12:00	00:12:03	Jiang	Speak fast. Otherwise, we couldn't finish it. We couldn't have lunch.	
235	00:12:03	00:12:04	Zhou	Isn't it this?	 #242. Tang (middle) and Zhou (left) laughed and looked at the paper. Jiang (right) looked at the paper while speaking.
236	00:12:04	00:12:05	Tang	I choose five.	
237	00:12:05	00:12:06	Zhou	The black dots must be here.	
238	00:12:06	00:12:06	Tang	I choose five.	
239	00:12:07	00:12:07	Zhou	Why?	
240	00:12:10	00:12:11	Tang	I am just guessing.	
241	00:12:13	00:12:14	Zhou	Whatever I speculate the last one must be this.	
242	00:12:13	00:12:14	Jiang	What the hell are you talking about?	

The thinking trajectory of this 25-turn excerpt could be expressed as {… (0,1,0), (1,1,1), (0,0,0), (0,1,0), (1,0,0), (1,1,0), (0,1,1), (1,1,0), (0,1,0), (1,1,1), (1,0,0), (0,1,1), (1,1,0), (1,0,0), (1,1,0), (1,0,0), (1,1,1), (1,1,1), (1,1,0), (1,1,1), (1,1,1), (1,1,0), (1,1,1), (1,1,0), (1,1,0) …}. It was divided into six subsequences (see Fig. 6).

The GTS scores in Excerpt 2 were RGTS = 25/6 = 4.17, PGTS = 20/6 = 3.33, and CGTS = 9/6 = 1.50. The reciprocity, productiveness, and constructiveness of the group thinking of G8 improved markedly from the pre-test to the post-test when solving a pair of equivalent puzzles. At post-test, the group interacted reciprocally for more than four consecutive turns on average, which was double their pre-test performance on the same measure. More than half of the talk tools they employed at post-test belonged to You-Talk, such as "agree," "disagree," "build on others," and "press for reasoning." In particular, there was a long thinking flow from Turn 107 to 119 after some fragmented thinking sequences, which led them to the right answer. These results indicate that the members of G8 were more willing to listen to each other and engage with each other's voices (rather than focused only on expressing their individual ideas) when solving the equivalent puzzle at post-test than at pre-test.

4.2.1.2. Summary of G8. Throughout the pre-test group thinking tasks, G8 engaged in sufficient discussion on simple puzzles and tried hard to solve challenging puzzles. They made guesses about some difficult items and ultimately gave up checking their answers as they were rushing to finish the task, especially when they saw that other groups had finished. Although we asked the students not to have leaders in their groups, Liu spontaneously took on the role of leader in G8 due to his academic advantage and confidence in solving the puzzles.

As indicated by the excerpt from the pre-test puzzle discussion, Liu behaved authoritatively and controlled the decision-making. Qi thought deeply about the puzzles and had many constructive interactions with Liu but nevertheless relied on Liu to make decisions. Li tried hard to contribute but tended to be ignored and distrusted by both Liu and Qi.

In contrast, throughout the post-test group thinking tasks, these three students tried to persuade each other by providing explanations and did not rush to finish the task. They did not simply make guesses about the answers to challenging items; rather, they made a concerted effort to think of an answer, which helped them to succeed in solving some challenging puzzles.

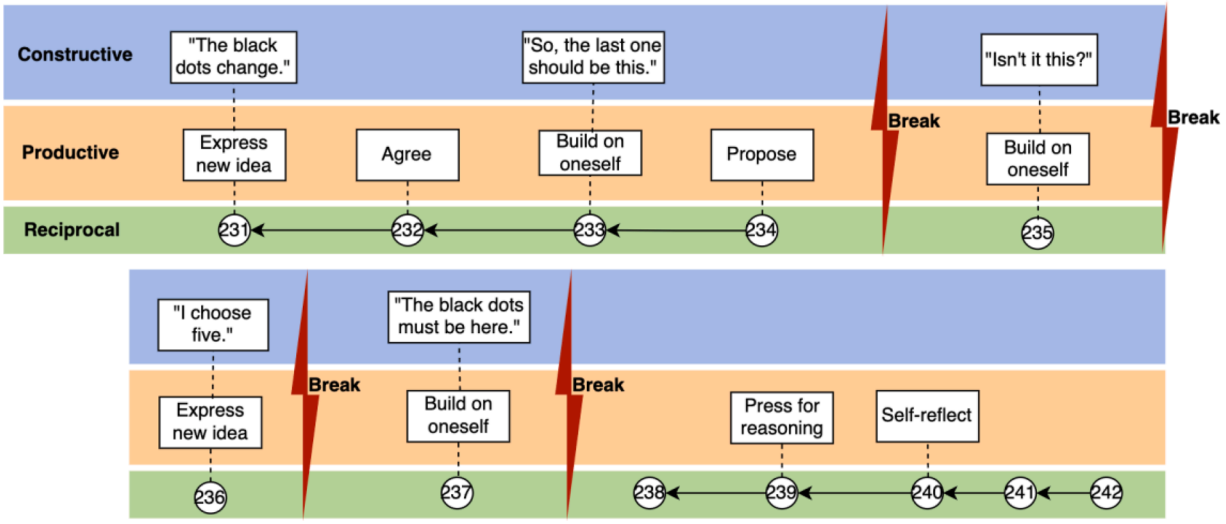


Fig. 7. Dynamics of GTS in G38 during the pre-test Puzzle 9.

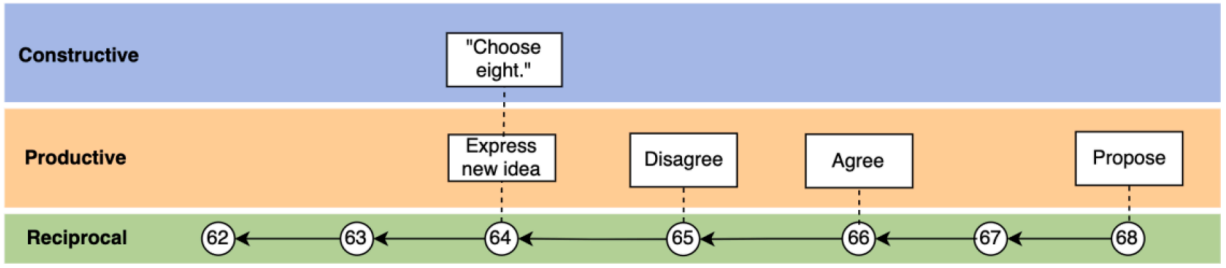


Fig. 8. Dynamics of GTS in G38 during the post-test Puzzle 10.

As indicated by the contrasting performance of G8 in the two puzzles, G8 developed a stronger awareness of the need to listen to and persuade others through explanation or elaboration. Liu maintained a leadership role but became less authoritative and more willing to explain to others and provide others with opportunities to speak. Li became more confident to speak out and insist on her ideas rather than just following others. She also received more respect and responses, although she was criticized by Liu and Qi for voicing an answer without explaining it. Qi had more opportunities to contribute her deep thinking. Together, G8 successfully solved the puzzle equivalent to that they had previously failed to solve due to a lack of authentic dialogue.

4.2.2. G38 in the comparison class

4.2.2.1. Comparison of performance on the selected pair of puzzles. When solving the challenging Puzzle 9 at pre-test, Zhou and Tang had constructive discussions about a solution. In Excerpt 3, they became a little stuck and Tang guessed an answer (#240). Jiang was not involved in solving the puzzle and only commented from time to time (#232, #234, #242), with these comments being ignored by Zhou and Tang. The group settled on an answer proposed by Zhou. Therefore, only Tang and Zhou contributed to the group's solution while Jiang was largely off-task and neglected. In addition, Tang tended to rely on Zhou for decision-making.

The thinking trajectory of this 12-turn excerpt could be expressed as {… (0,1,1), (1,1,0), (1,1,1), (1,1,0), (0,1,1), (0,1,1), (0,1,1), (0,0,0), (1,1,0), (1,1,0), (1,0,0), (1,0,0) …}. It was divided into five subsequences (see Fig. 7).

The GTS scores in Excerpt 3 were RGTS = 12/5 = 2.40, PGTS = 9/5 = 1.80, and CGTS = 5/5 = 1.00. The group thinking of G38 in this excerpt was relatively fragmented, with only 2.4 consecutive reciprocal turns on average. More than half of the talk tools the group used were I-Talk tools, such as "express new idea," "build on oneself," and "self-reflect." The group thinking flow was broken when Zhou focused on developing her own thinking (#235, #237) while Tang was trying to guess an answer (#236, #238).

When solving a similar puzzle (Puzzle 10) at post-test (see Excerpt 4), G38 again rushed to achieve a consensus and engaged in few constructive discussions. Jiang initially voiced an answer (#64). Zhou disagreed immediately but could not determine the correct answer (#65). Instead of waiting for her, Tang agreed with Jiang (#66) and suggested that Zhou should also trust Jiang's answer

(#68). The group then rushed on to the next problem even though Zhou was remained confused (#67).

Excerpt 4. "Let's listen to him."

Turn	Start time	End time	Speaker	Content	Embodied actions
62	0:04:31	0:04:31	Tang	This is…	
63	0:04:36	0:04:36	Zhou	This is…	
64	0:04:36	0:04:36	Jiang	Choose eight. No! It should be…	
65	0:04:37	0:04:40	Zhou	wait…	
66	0:04:39	0:04:40	Tang	Eight. It is eight.	
67	0:04:41	0:04:43	Zhou	If eight…	#64. Jiang (right) looked at the paper while speaking. Zhou (left) looked at the paper with a hand on the paper. Tang (middle) lay back in her chair and looked at the paper.
				Let's listen to him.	
➡**68**	0:04:47	0:04:49	Tang	Come on, next problem.	

The thinking trajectory of this seven-turn excerpt could be expressed as {… (0,0,0), (1,0,0), (1,1,1), (1,1,0), (1,1,0), (1,0,0), (1,1,0) …}. There was no break point in this thinking sequence (see Fig. 8).

The GTS scores in Excerpt 4 were RGTS = 7/1 = 7, PGTS = 4/1 = 4, and CGTS = 1/1 = 1. G38 engaged in seven consecutive turns of reciprocal discussion when working on this puzzle. They adopted four talk tools on one new idea and reached a consensus very efficiently. They resolved a disagreement by trusting Jiang, who was the most academically gifted, rather than by seeking explanations.

4.2.2.2. Summary of G38

Consistent with the self-reported background information (see Table 3), Jiang was the least talkative member of the group but was trusted by the other members, especially by Tang, because Jiang had the highest academic status. Tang was talkative but was likely to follow the other members in decision-making because she had the lowest academic status. She talked much with Zhou and acted as the facilitator of the group. Zhou contributed the most to puzzle-solving and was also trusted by Tang.

Jiang barely participated throughout the pre-test group thinking tasks. Zhou and Tang held a constructive but prolonged discussion at the beginning of the pre-test, which greatly slowed their progress. They made guesses on the last three challenging puzzles because they were rushing to finish the task—even though they were not out of time—as they had seen the other groups had handed in their answers. Zhou and Tang's guesses turned out to be wrong, which indicates that they may have achieved a higher score if they had maintained constructive discussions.

As indicated by the contrasting performance of G38 in the selected pair of puzzles, G38 engaged in significantly less constructive discussion at post-test than at pre-test as they were rushing to find the answer. They had gained experience in the type of puzzles exemplified by the pre-test thinking tasks. Moreover, they might have remembered some of the puzzles vaguely from the individual thinking test they had taken at pre-test. Therefore, they voiced their answers directly and quickly reached a consensus by directly trusting the more academically gifted members. It was obvious that Tang trusted Jiang, who was the most academically gifted in the group, and did not seek an explanation despite not understanding the answer. G38's rush to finish the task also made it less likely that they would have high-quality discussions than if they had not rushed.

5. Discussion

The present study revealed that there were no significant differences in GTM scores between the two classes, although there were more groups in the intervention class that showed an increase in their grouping values. The non-significant results suggest that it is still unclear whether or not our teaching-talk-for-thinking program can outperform implicit learning from repeated collaboration practice. Several potential explanations for this finding were proposed. Firstly, this finding may indicate the challenges associated with improving group thinking skills. Previous research, such as the Thinking Together project, reported significant individual thinking improvements after approximately a semester of explicit exploratory talk learning (Littleton & Mercer, 2013; Wegerif et al., 1999). However, there has been a lack of research on talk intervention programs that describe changes in group thinking (using the GTM criterion or other criteria).

Secondly, repeated collaboration practice may also facilitate better group thinking performance in tasks. Group thinking is related to individual social sensitivity and individual communication skills (Chen et al., 2012a,b; Woolley et al., 2010). The accumulation of collaboration experience may strengthen group members' knowledge of each other, allow them to practice their individual communication skills, and make them feel sufficiently psychologically safe to express their individual viewpoints. Group thinking may,

Table A1
Practice tasks developed in the study.

Label of the task	Description
Snake	Students need to estimate the number of stones that a snake would occupy when it straightened its body.
Goat and lion	Five people play either an honest goat or a lying lion in a game. Students need to figure out their roles based on their statements.
Candy	Students need to calculate the number of combinations to eat out six candies if only one or two candies can be eaten every time.
Car	Students need to draw satisfactory travel routes for a car to arrive at a destination without passing any road crossing more than once.
Ticket	Students need to design a ticket plan for three adults and 14 children that would cost them the least to visit a museum.
Numbers	Students need to fill numbers one to three to eight circles to make the sum the least.
Bridge	Students need to design a bridge-crossing plan for four people that would take the least amount of time.
Wall	Students need to tile a certain sized wall with six given tiles and draw ten different solutions.
Triangle	Students need to count the number of triangles in a complex geometrical pattern.

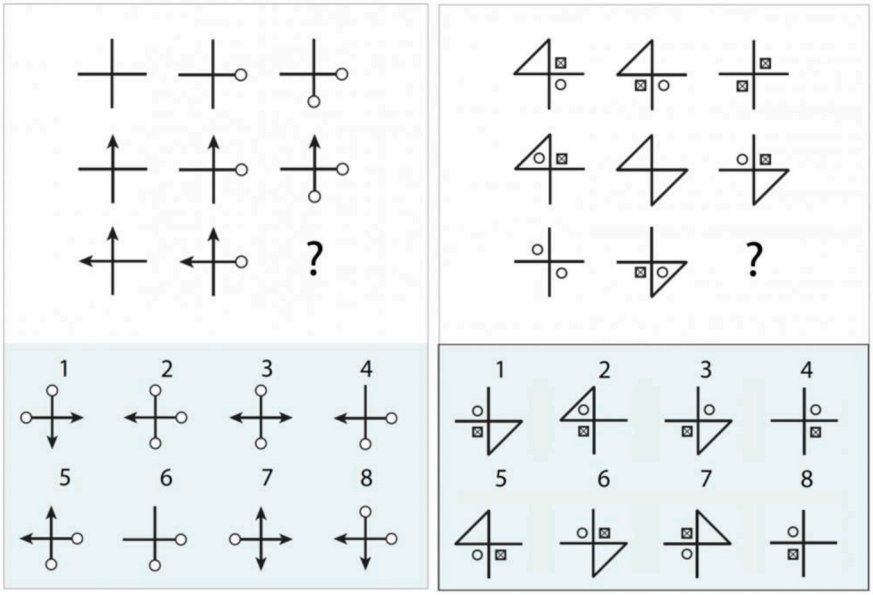


Fig. A1. Two example puzzles from the GTM (Wegerif et al., 2017).

therefore, be improved through the fulfillment of the potential of individual members.

Thirdly, the GTM test may be limited in measuring group thinking. The visual puzzles in GTM imitate the Raven's Standard Progressive Matrices which is used to test individual intelligence. Through our intervention, we discovered that the simple or difficult visual puzzles in the GTM test were not ideal for group discussions. Typically, students simply provided answers for the simple puzzles or guessed answers for the difficult ones. As a result, the GTM test may have limited ability to stimulate group thinking and consequently hinder its effectiveness in measuring the quality of group thinking.

Our analysis of the collaborative discourse in the group thinking tests unveiled varying patterns in the changes of talk tool usage between the two classes. Specifically, the comparison class showed a significant increase in the use of I-Talk at post-test compared to pre-test, while the intervention class showed a significantly greater use of You-Talk and We-Talk-Equity. This suggests that the repeated practice of collaboration improves group thinking in a different manner than the teaching-talk-for-thinking program. The high usage of I-Talk in the comparison class indicates that individual group members were more engaged in self-expression at post-test than at pre-test, leading to an increased degree of actualization of individual thinking and subsequently improved group thinking. Conversely, the intervention class improved group thinking primarily through the stronger interdependence of individuals.

A high usage of You-Talk tools signifies a strong engagement with the ideas of other group members, indicating a social and cognitive interdependence essential for group effectiveness (Miyake & Kirschner, 2014; Van Den Bossche et al., 2006). Stahl (2014) suggests that effective group cognition requires shared understanding amongst members. You-Talk tools, which involve building on each other's viewpoints and expressing agreement, are conducive to establishing and maintaining a shared understanding. Additionally, the use of We-Talk-Equity tools helped the intervention groups achieve more balanced interactions, which is crucial for effective collaboration (Asterhan & Schwarz, 2009; Dillenbourg et al., 2016), preventing information loss, dominance by a majority, and limiting a team's potential to solve problems (Borge & Carroll, 2014; Woolley et al., 2010).

The GTS measures showed that the intervention class exhibited significantly greater improvement in thinking sustainability than the comparison class. While the comparison class showed a high usage of I-Talk tools, this was accompanied by more fragmented group thinking flow. Hence, repeated collaboration practice alone appeared insufficient to bring about group-level changes, despite the fact

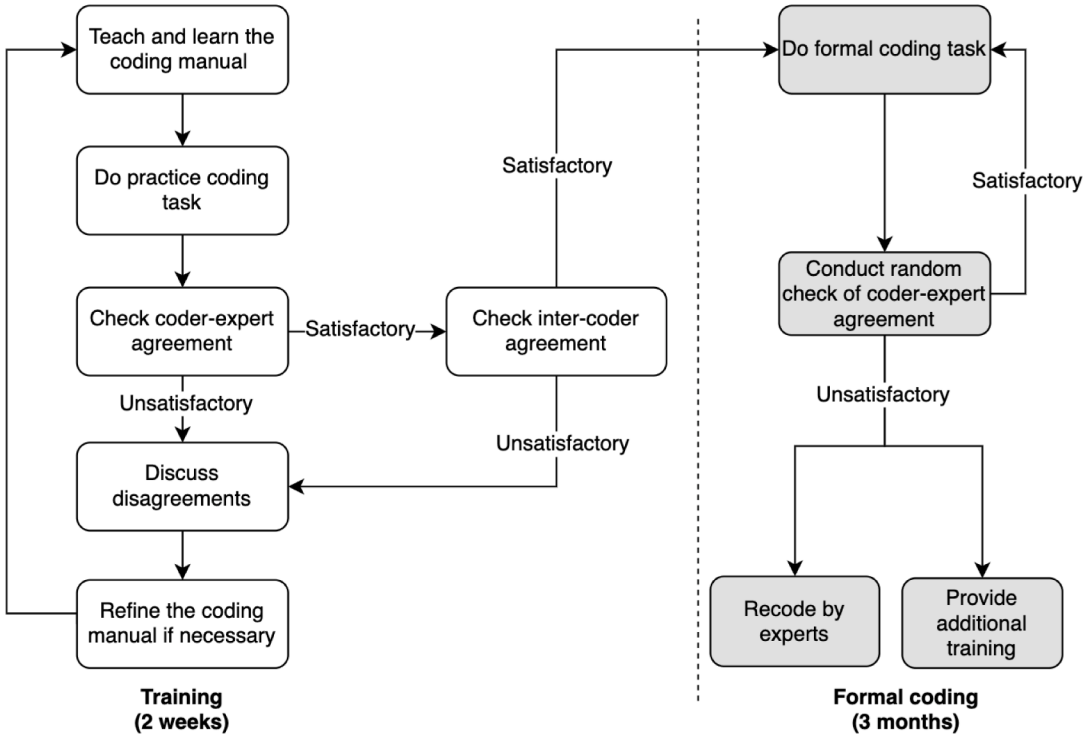


Fig. A2. A flow diagram on the training of qualified coders and the procedure of formal coding.

that it made individuals more comfortable with expressing their ideas. Conversely, the intervention class demonstrated longer and more productive reciprocal sequences, consistent with the higher usage rates of You-Talk and We-Talk-Equity tools. Long response sequences are also critical for effective group cognition (Stahl, 2014), enabling groups to engage in deep collective thinking and jointly solve problems. However, it is unclear why the significant discourse-level changes did not translate into significant changes in the thinking test scores in this study.

The qualitative case analysis of two successful groups further illustrate how the repeated collaboration practice and the teaching-talk-for-thinking program may improve group thinking in different ways. The groups in the intervention class spontaneously developed a culture of authentic dialogue. The group thinking scores of G8 in the intervention class were increased by the group members adopting a more open-minded approach than they had previously used, as this new approach enabled them to build and maintain shared understandings. The group thinking scores of G38 in the comparison class were increased mainly by this group's reliance on an academically gifted member and by the group improving its problem-solving strategies, such as its time management. The productive peer-talk intervention improved group thinking by helping individuals reach their potential and triggering the added value of grouping, whereas the practice-orientated approach improved group thinking by leading to the spontaneous adjustment of obvious collaboration problems and the improvement of problem-solving strategies.

6. Limitations and suggestions for future research

There are several limitations to this study. First, there was a possible practice effect from the pre- to the post- thinking tests. This was because the GTM just contains two identical thinking tests, which meant that a practice effect could not be excluded from the pre–post design. We attempted to minimize the practice effect by switching the test sets for individuals and groups in the pre- and post-tests. However, some students appeared to still remember the items even after three months; some of these students had sufficient familiarity with the questions to solve the problems faster than if they had not seen the problems, while a few students rapidly retrieved an answer from their memory. The results also showed that the groups in both the intervention and comparison classes exhibited significantly fewer turns at post-test than at pre-test. This suggests there is a need to be cognizant of the practice effect when using the GTM to measure the effects of an intervention on group thinking with a pre–post design. Moreover, future studies should explore how the practice effect may affect test scores, and improvements may be needed in the GTM to better suit the needs of a pre–post design. For example, additional matching test item sets may need to be constructed to eliminate the practice effect.

Second, the design of the comparison class incorporated some evaluative feedback on dialogue, which made the classroom collaborations not purely natural. As it is not clear to what extent this simple evaluative feedback on the quality of dialogue influences collaborations, future studies could incorporate another comparison class that engages only in practice without any feedback. This would allow examination of whether students spontaneously improve their individual thinking skills without the pressure of

evaluating their dialogue and whether group thinking skills may spontaneously improve under these conditions.

Third, different groups progressed in different ways. Although there were more groups in the intervention class than in the comparison class that markedly improved their group thinking skills, there were also four groups in the intervention class that markedly decreased their group thinking skills after the intervention. It remains unclear why these groups did not benefit from the intervention and what additional scaffolding should be provided for such groups. It also remains unclear whether the group thinking skills of these groups declined in the same way as that of groups in the comparison class. If such marked changes in group thinking skills in a non-negligible proportion (25%) of groups were not due to some characteristic group features, this may indicate the GTM lacks validity as a one-time test for measuring group thinking.

7. Conclusion

This study introduced a technology-advanced teaching-talk-for-thinking program aimed at explicitly teaching students fine-grained talk moves as tools for achieving equity, open-mindedness, and convergence in dialogic collaborative problem-solving. In comparison to a commonly used approach in classrooms that provides students with ample opportunities to collaborate but offers no guidance on peer talk, this program proved more beneficial in engaging students verbally and promoting persistence in group thinking, though it did not result in higher scores in the group thinking tests. The study thus demonstrated the impact of explicitly teaching productive peer talk on group thinking, while also raising the question of whether learning talk could better improve group thinking skills than accumulating collaboration experience. This research makes both theoretical and empirical contributions to our understanding of group thinking and calls for further enquiry into this area, as well as continued improvements to teaching-talk-for-thinking intervention programs.

Compliance with ethical standards

Conflict of interest

The authors have no relevant financial or non-financial interests to disclose.

Research involving in human rights

The procedures that involved human participants adhered to the ethical standards set by the Human Research Ethics Committee (HREC) of the University of Hong Kong (reference number: EA1903032).

Informed consent

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

CRediT authorship contribution statement

Liru Hu: Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft. **Gaowei Chen:** Supervision, Writing – review & editing, Validation. **Jiajun Wu:** Data curation, Software, Visualization, Validation.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by International Research Institute for the Learning Sciences (#EDT/2020/1/1) and Hong Kong Research Grants Council, University Grants Committee (Grant No.: 17605221).

Appendix

Fig. A1, Fig. A2

References

Agar, M. (2006). An ethnography by any other name …. *Forum Qualitative Sozialforschung /Forum: Qualitative Social Research, 7*(4). http://nbn-resolving.de/urn:nbn:de:0114-fqs0604367.

Alexander, R. (2005). Education, culture and cognition: Intervening for growth international association for cognitive education and psychology (IACEP). In *10th International Conference.* University of Durham.

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

Australian Math Trust. (n.d.). *Australian mathematics competition.* Retrieved January 22, 2021, from https://www.amt.edu.au/australian-mathematics-competition.

Awang, Z. H. (2015). Validating the measurement model: CFA. Ed.. In Z. H. Awang (Ed.), *A handbook on sem* (2nd ed, pp. 54–73). Kuala Lumpur: Universiti Sultan Zainal Abidin
Baker, M. J., Andriessen, J., & Schwarz, B. B. (2020). Collaborative argumentation-based learning. In N. Mercer, R. Wegerif, & L. Major (Eds.), *The Routledge international handbook of research on dialogic education* (pp. 76–88). Routledge.
Bakhtin, M. M. (1929/1984). *Problems of Dostoevsky's poetics (C. Emerson, Ed., Trans.). Manchester University Press*. Manchester University Press.
Bakhtin, M. M. (1981). *The dialogic imagination*. Austin: University of Texas Press.
Bedford, A. (1997). On Clark–Watson\'s tripartite model of anxiety and depression. *Psychological Reports, 80*(1), 125–126.
Bialystok, E., Craik, F., Grady, C., Chau, W., Ishii, R., Gunji, A., et al. (2005). Effect of bilingualism on cognitive control in the Simon task: Evidence from MEG. *NeuroImage, 24*, 40–49. https://doi.org/10.1016/j.neuroimage.2004.09.044
Borge, M., & Carroll, J. M. (2014). Verbal equity, cognitive specialization, and performance. In *Proceedings of the 18th international conference on supporting group work* (pp. 215–225).
Bridges, S. M., Hmelo-Silver, C. E., Chan, L. K., Green, J. L., & Saleh, A. (2020). Dialogic intervisualizing in multimodal inquiry. *International Journal of Computer-Supported Collaborative Learning, 15*(3), 283–318. https://doi.org/10.1007/s11412-020-09328-0
Chen, G., Chiu, M. M., & Wang, Z. (2012). Predicting social cues during online discussions: Effects of evaluations and knowledge content. *Computers in Human Behavior, 28*(4), 1497–1509.
Chen, G., Chiu, M. M., & Wang, Z. (2012). Social metacognition and the creation of correct, new ideas: A statistical discourse analysis of online mathematics discussions. *Computers in Human Behavior, 28*(3), 868–880. https://doi.org/10.1016/j.chb.2011.12.006
Chiu, M. M., & Khoo, L. (2003). Rudeness and status effects during group problem solving: do they bias evaluations and reduce the likelihood of correct solutions? *Journal of Educational Psychology, 95*(3), 506–523. https://doi.org/10.1037/0022-0663.95.3.506.
Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/bf02310555
Davis, M. H. (1983). A mulitdimensional approach to individual differences in empathy. *Journal of Personality and Social Psychology, 44*(1), 113–126. https://doi.org/10.1037/0022-3514.44.1.113
Dillenbourg, P., Lemaignan, S., Sangin, M., Nova, N., & Molinari, G. (2016). The symmetry of partner modelling. *International Journal of Computer-Supported Collaborative Learning, 11*(2), 227–253. https://doi.org/10.1007/s11412-016-9235-5
Dressman, M. (2004). Dewey and Bakhtin in dialogue: From Rosenblatt to a pedagogy of literature as social, aesthetic practice. *Bakhtinian perspectives on language, literacy, and learning, 1*, 34–52.
Faul, F., Erdfelder, E., Lang, A.-. G., & Buchner, A (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146
Fujita, T., Doney, J., Flanagan, R., & Wegerif, R. (2019). Collaborative group work in mathematics in the UK and Japan: Use of group thinking measure tests. *Education 3–13, 49*(2), 119–133. https://doi.org/10.1080/03004279.2019.1701513
Gillies, R. M. (2019). Promoting academically productive student dialogue during collaborative learning. *International Journal of Educational Research, 97*, 200–209. https://doi.org/10.1016/j.ijer.2017.07.014
Green, J. L., & Bridges, S. M. (2018). Interactional ethnography. Eds.. In F. Fischer, C. E. Hmelo-Silver, S. R. Goldman, & P. Reimann (Eds.), *International handbook of the learning sciences* (pp. 475–488). New York: Routledge
Hinton, P. R., McMurray, I., & Brownlow, C. (2014). *SPSS explained*. London: Routledge.
Hooper, M., Mullis, I., & Martin, M. (2013). TIMSS 2015 context questionnaire framework, 2013. In I. V. S. Mullis, & M. O. Martin (Eds.), *TIMSS 2015 assessment frameworks* (Eds., pp. 61–82). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
Hu, L., & Chen, G. (2021). Towards a complex systems perspective on the temporal patterns of dialogic collaborative problem solving. *Frontiers in Psychology, 12*, 735534. https://doi.org/10.3389/fpsyg.2021.735534.
Hu, L., & Chen, G. (2022). Exploring turn-taking patterns during dialogic collaborative problem solving. *Instructional Science, 50*(1), 63–88. https://doi.org/10.1007/s11251-021-09565-2.
Hu, L., & Chen, G. (2023). A systematic review and meta-analysis of productive peer talk moves. *Journal of Behavioral Education*, 1–33. https://doi.org/10.1007/s10864-023-09513-9.
Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–31. https://doi.org/10.1080/10705519909540118
Hu, L., Wu, J., & Chen, G. (2022). iTalk–iSee: A participatory visual learning analytical tool for productive peer talk. *International Journal of Computer-Supported Collaborative Learning*, 1–29.
Hu, L. (2021). Conceptualization and operationalization of group thinking sustainability in dialogic collaborative problem solving. *Thinking Skills and Creativity, 42*, Article 100964. https://doi.org/10.1016/j.tsc.2021.100964. September.
International Mathematics Assessments for Schools. (n.d.). *About: International mathematics assessment for schools (IMAS)*. Retrieved January 22, 2021, from https://math4schools.wordpress.com/about/.
King, A. (1997). ASK to THINK-TEL WHY: A model of transactive peer tutoring for scaffolding higher level complex learning. *Educational Psychologist, 32*(4), 221–235.
King, A. (2008). Structuring peer interaction to promote higher-order thinking and complex learning in cooperating groups. In R. Gillies, A. Ashman, & J. Terwel (Eds.), *The teacher\'s role in implementing cooperative learning in the classroom* (Eds., pp. 73–91). New York: Springer.
Kirilenko, A. P., & Stepchenkova, S. (2016). Inter-coder agreement in one-to-many classification: Fuzzy kappa. *PloS one, 11*(3), 1–15. https://doi.org/10.1371/journal.pone.0149787
Kolikant, Y. B., & Pollack, S. (2020). The power of a dialogical framework to articulate collaborative learning in the 21st century. Eds.. In N. Mercer, R. Wegerif, & L. Major (Eds.), *The routledge international handbook of research on dialogic education* (pp. 634–646). Routledge
La Greca, A. M., & Stone, W. L. (1993). Social anxiety scale for children-revised: Factor structure and concurrent validity. *Journal of Clinical Child Psychology, 22*(1), 17–27.
Littleton, K., & Mercer, N. (2013). *Interthinking: Putting talk to work*. Routledge.
Matusov, E., Marjanovic-Shane, A., & Gradovski, M. (2019). *Dialogic pedagogy and polyphonic research art: Bakhtin by and for educators*. New York: Springer. https://doi.org/10.1057/978-1-137-58057-3
Mayer, R. E. (2014). Principles based on social cues in multimedia learning: Personalization, voice, image, and embodiment principles. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 345–368). Cambridge University Press.
Mercer, N., Wegerif, R., & Dawes, L. (1999). Children\'s talk and the development of reasoning in the classroom. *British Educational Research Journal, 25*(1), 95–111. https://doi.org/10.1080/0141192990250107
Mercer, N. (2002). *Words and minds: How we use language to think together*. Routledge.
Mercer, N. (2013). The social brain, language, and goal-directed collective thinking: A social conception of cognition and its implications for understanding how we think, teach, and learn. *Educational Psychologist, 48*(3), 148–168. https://doi.org/10.1080/00461520.2013.804394
Miyake, N., & Kirschner, P. A. (2014). The social and interactive dimensions of collaborative learning. Ed.. In R. Keith Sawyer (Ed.), *The cambridge handbook of the learning sciences* (2nd Ed., pp. 418–438). Cambridge University Press. https://doi.org/10.1017/CBO9781139519526.026
Noroozi, O., Teasley, S. D., Biemans, H. J., Weinberger, A., & Mulder, M. (2013). Facilitating learning in multidisciplinary groups with transactive CSCL scripts. *International Journal of Computer-Supported Collaborative Learning, 8*(2), 189–223.
Resnick, L. B., Michaels, S., & O'Connor, C. (2010). How (well structured) talk builds the mind. In D. Preiss, & R. Sternberg (Eds.), *Innovations in educational psychology: Perspectives on learning, teaching and human development* (pp. 163–194). New York, NY: Springer.
Rojas-Drummod, S., Pérez, V., Vélez, M., Gómez, L., & Mendoza, A. (2003). Talking for reasoning among Mexican primary school children. *Learning and Instruction, 13*(6), 653–670. https://doi.org/10.1016/S0959-4752(03)00003-3
Simms, A., & Nichols, T. (2014). Social loafing: A review of the literature. *Journal of Management Policy and Practice, 15*(1), 58–67.

Stahl, G. (2014). The constitution of group cognition. Ed.. In L. Shapiro (Ed.), *The routledge handbook of embodied cognition* (pp. 335–346). Routledge
Stahl, G. (2016). From intersubjectivity to group cognition. *Computer Supported Cooperative Work, 25*, 355–384. https://doi.org/10.1007/s10606-016-9243-z
TIMSS & PIRLS International Study Center (2015). *TIMSS 2015 item information tables–fourth grade*. Retrieved from https://timssandpirls.bc.edu/timss2015/international-database/downloads/T15_G4_ItemInformation.zip.
Taber, K. S. (2018). The use of Cronbach\'s alpha when developing and reporting research instruments in science education. *Research in science education, 48*(6), 1273–1296.
Topping, K. J., & Trickey, S. (2013). The role of dialog in philosophy for children. *International Journal of Educational Research, 63*, 69–78. https://doi.org/10.1016/j.ijer.2013.01.002
Trausan-Matu, S., Wegerif, R., & Major, L. (2021). Dialogism. Eds.. In U. Cress, J. Oshima, C. Rosé, & A. Wise (Eds.), *International handbook of computer-supported collaborative learning* (pp. 219–239). Springer International Publishing
Van Den Bossche, P., Gijselaers, W. H., Segers, M., & Kirschner, P. A. (2006). Social and cognitive factors driving teamwork in collaborative learning environments: Team learning beliefs and behaviors. *Small Group Research, 37*(5), 490–521. https://doi.org/10.1177/1046496406292938
Vass, E., Littleton, K., & Littleton, K. (2010). Peer collaboration and learning in the classroom. Eds.. In C. Wood, & J. Kleine Staarman (Eds.), *International handbook of psychology in education* (pp. 105–136). Leeds, UK: Emerald
Vygotsky, L. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.
Webb, N. M., Franke, M. L., Ing, M., Wong, J., Fernandez, C. H., Shin, N., et al. (2014). Engaging with others' mathematical ideas: Interrelationships among student participation, teachers' instructional practices, and learning. *International Journal of Educational Research, 63*, 79–93. https://doi.org/10.1016/j.ijer.2013.02.001
Wegerif, R., Mercer, N., & Dawes, L. (1999). From social interaction to individual reasoning: An empirical investigation of a possible sociocultural model of cognitive development. *Learning and Instruction*. https://doi.org/10.1016/S0959-4752(99)00013-4
Wegerif, R., Fujita, T., Doney, J., Perez Linares, J., Richards, A., & van Rhyn, C. (2017). Developing and trialing a measure of group thinking. *Learning and Instruction, 48*, 40–50. https://doi.org/10.1016/j.learninstruc.2016.08.001
Wegerif, R. (2013). Dialogic: Education for the internet age. *London: Routledge*. https://doi.org/10.4324/9780203111222
Wertsch, J. V. (1991). *Voices of the mind: A sociocultural approach to mediated action*. Cambridge, MA: Harvard University Press.
Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science, 330*(6004), 686–688. https://doi.org/10.1126/science.1193147
Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist, 34*(6), 806–838. https://doi.org/10.1177/0011000006288127
Wright, D. B. (2006). Comparing groups in a before–after design: When t-test and ANCOVA produce different results. *British Journal of Educational Psychology, 76*(3), 663–675. https://doi.org/10.1348/000709905X52210
Xie, K., Miller, N. C., & Allison, J. R. (2013). Toward a social conflict evolution model: Examining the adverse power of conflictual social interaction in online learning. *Computers and Education, 63*, 404–415. https://doi.org/10.1016/j.compedu.2013.01.003.
Zhu, H., & Sun, W. (2018). *Solution of Po Leung Kuk primary mathematics world contests* (2nd edition). Beijing: Science Press.