Learning, Culture and Social Interaction 40 (2023) 100717

Contents lists available at ScienceDirect

Learning, Culture and Social Interaction

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

Full length article

In pursuit of a more unified method to measuring classroom dialogue: The dialogue elements to compound constructs approach

Edith Bouton [*], Christa S.C. Asterhan

The Seymour Fox School of Education, Hebrew University of Jerusalem, Israel

ARTICLE INFO

Keywords:
Coding classroom dialogue
Academically productive dialogue
Quantitative research
Inter-rater reliability

ABSTRACT

There is increasing scholarly agreement about the key features of academically productive classroom dialogue. Yet, despite this emerging conceptual consensus, the ways in which it is measured and coded in quantitative research efforts vary significantly. In order to communicate, compare and integrate findings from this rich body of empirical research and to further theory refinement, a more unified approach to measuring classroom dialogue is needed. We selected seven well-known coding frameworks and identified a set of nine particle-like dialogue elements (DEs), that lie at the basis of different coding categories, appear frequently in classroom dialogue, and can be reliably coded at the conversational turn level. We then demonstrate how a much larger set of "compound" dialogue constructs can be recreated post-coding, by flagging co-occurrences of different DEs and accounting for the majority of coding categories in each of the seven frameworks. This Dialogue Elements to Compound Constructs Approach (DECCA) then enables interrater reliability, while simultaneously maintaining the flexibility and comprehensiveness needed to enable quantitative research on a large variety of research questions with a single methodological approach. The implications for future research and theory are discussed.

1. Introduction

Research on academically productive dialogue (APD hereafter) for learning can be traced back to the 1970s, but has really burgeoned in the 21st century. These efforts have been multidisciplinary in nature, characterized by complementary research paradigms and topical foci, and have accumulated into a rich and varied body of theoretical knowledge and evidence (e.g., Alexander, 2018; Kershner et al., 2020; Littleton & Mercer, 2013; Mercer & Littleton, 2007; Resnick et al., 2015). Reviews of this multidisciplinary body of work have highlighted the emergence of a conceptual consensus about the type of classroom talk that promotes student learning and cognitive development (e.g., Asterhan and Schwarz, 2016; Howe & Abedin, 2013; Howe & Mercer, 2016; Kim & Wilkinson, 2019; Resnick et al., 2018a, 2018b): In APD, teachers engage their students with open-ended, authentic questions or tasks and encourage them to share their ideas and thinking processes, to build on each other's contributions, to (critically) explore alternative viewpoints, and to use external knowledge sources to evaluate them. This form of classroom dialogue has been described and identified across multiple countries and cultures, multiple learning settings (i.e., teacher-led classrooms, peer group work, higher, elementary, and secondary education, parent-child interactions, and so on), multiple communication formats (e.g., online, offline, textual, and oral),

* Corresponding author at: Hebrew University of Jerusalem, Mt Scopus 9190501, Israel.
E-mail address: edith.bouton@mail.huji.ac.il (E. Bouton).

https://doi.org/10.1016/j.lcsi.2023.100717
Received 24 April 2022; Received in revised form 2 April 2023; Accepted 13 April 2023
Available online 11 May 2023
2210-6561/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

and across multiple disciplinary subjects (such as philosophy, mathematics, science, history and language arts) (Resnick et al., 2015).

Yet, this emerging conceptual consensus about APD (i.e., the theoretical definitions) is not reflected in the ways it has been translated into measurable variables (i.e., the operational definitions). The ways in which APD is measured vary significantly across settings, research teams, and studies (Hennessy et al., 2020; Kim & Wilkinson, 2019; Song et al., 2019). In quantitative research efforts on classroom dialogue, the main tool for measurement is the systematic categorization of selected aspects of the observed interactions, that is: coding (Hennessy, 2020; Shaffer & Ruis, 2020). Admittedly, when the object of inquiry is as multi-faceted and dynamic as human, multi-participant discourse, a certain extent of variance in assessment tools may be expected. Yet, the plethora of coding tools available in empirical APD research seems disproportional and unnecessary, especially in light of the emerging conceptual consensus.

We posit that this lack of consensus on measurement stymies progress in the empirical study of APD as it obstructs our ability to compare, replicate and refine research findings from quantitative research in further detail (see also Pauli & Reusser, 2015; Tschan et al., 2018). This becomes even more challenging when empirical research reports do not include sufficient and transparent information on the rationale, development process, and inter-rater reliability rates of the coding scheme used, nor a systematic comparison with existing tools.

Let us consider, for example, the issue of identifying the exact classroom dialogue elements that produce academic gains. The existing research literature provides different answers to this seemingly basic question. For example, in some studies, explanations and elaborations have emerged as predictors of student gains (e.g., Chi & Menekse, 2015; Gillies, 2015; Webb et al., 2009, 2014), whereas in others, only disagreements and challenges, but not consensual explanations, were associated with learning (e.g., Asterhan and Schwarz, 2009; Howe, 2009; Larrain et al., 2019). In a more recent study, Howe et al. (2019), found that only the *combination* of queries with student-produced elaborations was associated with student achievement gains. Other findings highlight the importance of open-ended teacher questions (Nystrand et al., 2003) or of critical questions (Nussbaum, 2011).

Each of these dialogue moves (elaborations, challenges, critical questions, open-ended questions) is readily recognized as a part of the general theoretical construct of APD. The fact that different aspects of APD emerged as predictors of learning gains in different studies could potentially be explained by differences in the settings in which they were situated. Indeed, it is probable that the prominence of a particular APD element over others is contingent on factors such as culture, task setting, age group, and/or the student outcome targeted. However, in order to make systematic progress on this front and explore such contingencies, comparable measurement tools are necessary. Without them, any difference in outcomes across studies could be an artifact of the different ways in which APD was measured and coded.

In the next section, we consider three central dimensions along which different existing coding tools can be characterized: The number of coding categories included (representing issues of reliability and comprehensiveness), which speakers are targeted (teachers or students), and what type of talk moves are included (APD talk moves only, or less productive talk moves as well). As we will show, each of these three dimensions represents a tension between two seemingly opposing demands of measurement of APD, which has led to compromises and preferring one characteristic over the other. In this work, we argue that, ideally, such compromises should and do not need to be made. Building on the accumulative insights and progress that has been made in the field thus far, we aim to develop a coding approach that will satisfy all six demands at once.

1.1. Stability and consistency vs. sensitivity and comprehensiveness

One dimension on which existing coding tools vary, is the number of distinctive dialogue features they each aim to capture (i.e., the number of coding categories). To obtain a comprehensive representation of classroom talk that captures minute and nuanced differences between types of dialogue moves, a large palette of coding categories is needed. However, working with many coding categories has several disadvantages: First, detecting and correctly applying the appropriate coding category to discern subtle differences require extensive coder training and expertise. Moreover, as the number of coding categories in a given tool increases, the prevalence of the particular dialogue feature that a single category is intended to capture diminishes. In fact, some of the dialogue moves that APD scholars are particularly interested in are rather infrequent in everyday classroom dialogue. Examples of such rarely occurring instances are, among others, a student contributing a reasoned rebuttal to a counterargument, or a teacher re-voicing a student's contribution. Despite their rarity, APD scholars are keen in documenting such talk moves when they do appear, as they are regarded as indicative of particularly productive, high-quality classroom dialogue (Asterhan & Schwarz, 2007, 2009; Kuhn & Udell, 2003; O'Connor & Michaels, 1993).

Unfortunately, both these concerns (variance in coding expertise, as well as the rarity of some dialogue features) are considerable hurdles for achieving the interrater reliability thresholds required by common academic standards (Hennessy, 2020; Hennessy et al., 2020). It is perhaps not surprising then that oftentimes interrater reliability measures are not at all, or only ambiguously reported (see Thurlings & den Brok, 2017, for similar arguments). In other cases, when reliability thresholds could not be met, due to the rarity of the nuanced dialogic constructs a coding category had intended to capture, researchers have been forced to omit them from analyses or to cluster them together in coarser, more inclusive categories (e.g., Chiu, 2015; Hennessy et al., 2020). Either way, the intended aim of providing a comprehensive and sensitive account of classroom dialogue is forfeited, as analyses can only be conducted on the more common and/or coarser dialogue features. In addition, most of these considerations are left out of the final report in the published papers, obscuring the transparency of the research process.

Limiting the number of coding categories is then a necessary requirement for achieving acceptable inter-judge reliability indices for classroom coding tools. It is not a sufficient one, however, and it also comes with a price tag: Intentionally keeping the number of variables down to a minimum requires generalization (i.e., disregarding differences between similar, but not identical constructs) which could compromise the discriminating qualities of the tool and its validity (Boyd & Markarian, 2015; Lefstein & Barak, 2020;

Mercer, 2010). If a coding tool is only capable of capturing coarse and frequent dialogue features (for example, "questions" or "reasoning" in general), could it be sensitive enough to capture the subtle and gradual, yet important changes in dialogue quality? For example, let us consider two rather distinctive instances of the general category *reasoning*: (1) a student providing a reasoned rebuttal to a counterargument that was proposed against his/her initial claim; and (2) a student responding to a teacher's request to explain why he/she thinks that a particular shape is a rectangle. Both student contributions would be considered instances of reasoning in the general sense, yet the former represents a higher level of sophistication.

Claims to insufficient sensitivity of existing measuring tools is at the heart of theoretical and practical critiques of quantitative research methods for exploring rich, complex and heavily situated phenomena such as classroom dialogue (see a brief summary of the main points in Lefstein & Barak, 2020). It could also be one of the reasons behind some curious, and theoretically inconsistent findings in quantitative APD research: For example, some large scale APD intervention projects have reported on significant improvements in individual student achievement scores gains, even though no significant improvement in the classroom dialogue's quality was found (Osborne et al., 2013; Reznitskaya et al., 2007). Another example comes from a long-term, comprehensive intervention program aimed to improve students' argumentation competencies: Whereas no improvements could be detected in the first year, they were found in the second and third year of the program (Kuhn & Zillmer, 2015). Finally, several studies have documented how classroom discourse quality increases and reaches a peak in the first stages of an intervention, but is followed by a decline in the later stages (Crowell & Kuhn, 2014; Osborne et al., 2019).

In all the aforementioned cases, it is possible that subtler, smaller refinements in dialogue did occur, yet the analytical tools used in these studies were not sensitive enough to detect them. To (dis)prove this conjecture, coding tools are needed that are not only reliable, but also sensitive enough to pick up on subtle changes.

In sum, the two criteria for good measurement of classroom dialogue, sensitivity and comprehensiveness, on the one hand, and stability and consistency, on the other, seem to be in direct conflict with one another. In the present work, we explore whether it is possible to find a solution in which both demands could be met, without comprising either.

1.2. Coding teacher talk or student talk?

A second dimension on which existing coding tools can be characterized is whether they are designed for coding teacher talk only, student talk only, or both (see also Hennessy, 2020). The choice to measure solely teacher or student talk appears to be a byproduct of different foci in research questions, settings and traditions. Scholars of peer collaborative learning, for example, have primarily focused on the role of dialogue in peer-directed learning settings, such as group work, peer tutoring and dyadic tasks (e.g., Asterhan & Schwarz, 2009; Chi & Menekse, 2015; Webb, 1989). In this research tradition, the extent and quality of student participation are considered the main catalyzer of learning gains. Even when peer interactions are recorded with teachers present, circulating between and monitoring peer-to-peer work, coding efforts mostly (or solely) focus on capturing student talk. Information about teachers and teacher behavior is often relegated to the section providing background descriptions.

Scholars of teacher-led, whole classroom dialogue, on the other hand, have traditionally focused on teacher talk moves (Correnti et al., 2015, 2021; Michaels et al., 2008). This can be attributed to several reasons: First of all, in dialogic settings, teachers are in a position of power and must take on the responsibility of speaking up, navigating and maintaining the conversation. This motivates scholars to focus on the teacher talk moves. Second, teacher talk can be transcribed and coded more readily than student talk in whole classroom dialogue: Teacher moves for encouraging dialogue often belong to a known repertoire, especially when the data is collected in the context of targeted, APD-focused intervention programs. Student contributions, on the other hand, are often more rushed and incomplete, as well as less structured and unexpected, and therefore harder to code with pre-defined categories meant to capture distinctive speech acts or talk moves. Also, teacher contributions are often more audibly clear, as they are used to speaking in front of an audience and recording equipment is often positioned in their proximity.

Third, it is a longstanding tradition in classroom research to correlate changes in teacher behavior (either following professional development interventions, or not) with student outcomes, as measured by achievement tests or self-report surveys. The underlying (implicit) assumption is that the impetus for improving academic outcomes is a change in teaching practices, at the teacher level. Therefore, even though these improved student outcomes are assumed to be mediated by changes in quality of student engagement, such processes are nevertheless not directly measured.

There are considerable advantages to using a single coding system for capturing both teacher and student contributions with one single set of coding categories: First, it opens the possibility of merging and comparing research findings from the two aforementioned research venues (i.e., student peer-to-peer talk and teacher-led classroom dialogue). Second, it would enable empirical explorations into the interface between teacher behavior and student participation in dialogue, whether during whole classroom or small group discussions. These include, but are not limited to, the following research questions: How, if at all, do students notice, choose, adopt, and adapt teacher talk moves? Are the teacher's discourse behaviors taken up and imitated by students in peer group behavior (Gillies, 2015; Webb et al., 2014)? How do different teacher moves elicit particular patterns of student participation? For these reasons, we posit that, ideally, a coding tool would contain one single set of coding categories for capturing both student and teacher talk.

1.3. Coding only productive talk moves or traditional classroom talk?

A third dimension on which classroom discourse assessment tools vary is whether they assess the appearance of typical APD moves, or whether they (also) include coding categories for other talk moves, that are not part and parcel to APD, or may even be considered antithetical to it. As with the previous dimension, the source for this variance also seems to stem from different research traditions. On the one hand, sociolinguistic research into classroom discourse often tends to focus on the documentation of typical and prevailing discourse patterns. Among others, it has shown that classroom discourse is often dominated by recurring cycles of initiation-response-evaluation (or feedback), which is commonly known as an IRE (or IRF) cycle (Coulthard, 1992; Galton et al., 1999; Mehan & Cazden, 2015): The teacher *Invites* students to respond to a closed question, which has a limited range of permissible answers. A student *Responds* with a brief and unelaborated answer, after which the teacher *Evaluates* it by saying (or implying) whether it is correct or not (outcome feedback), before launching a new question. The focus in this research tradition is on the documentation of existing interaction patterns, rather than on changing them, while speculating and reasoning about how these shape (and limit) student opportunities for learning. The common assumption seems to be that IRE-based classroom discourse is less productive for student learning (Galton et al., 1999; Howe & Abedin, 2013). Even though these are reasonable expectations, to the best of our knowledge, they are rarely directly tested.

In contrast, studies on the impact of targeted intervention programs to change classroom discourse tend to focus (either solely or predominantly) on the type of talk moves that the intervention sought to cultivate, that is: APD moves (Fishman et al., 2017; Kuhn & Zillmer, 2015). The rationale behind this approach seems to be that increased frequencies of typical APD moves is proof of the intervention's success. More traditional talk features, such as IRE, are often not documented, presumably because they are expected to be replaced by APD moves, and/or because they are not expected to benefit student learning anyway.

Yet, in order to test these assumptions empirically, it is necessary to measure both types of talk moves and to test the relation between them, as well as with student learning measures. Ideally, a single coding scheme should then include categories for both APD moves, as well as more traditional classroom talk moves.

1.4. The present study

In the present work, we aim to build on the emerging conceptual consensus concerning APD, as well as on the present state of the art in APD coding methodologies, to develop a coding approach that is able to meet all of the six aforementioned demands: A coding tool that is comprehensive and sensitive to subtle differences and a wide variety of classroom discourse features, yet reliable and consistent in its measurement. A tool that captures both student and teacher talk moves, and captures academically productive, as well as more traditional classroom talk features.

To achieve this goal, we assembled a set of well-established, existing APD coding frameworks that have been developed and used in quantitative research efforts by leading scholars in the field. We then explored whether it is possible to identify a small set of recurring, basic elements of classroom dialogue across and within different frameworks. Our efforts to uncover basic, elementary similarities across a large set of different coding categories from a varied set of coding schemes are in many ways analogous to those of a chemist breaking down a large and varied group of compound substances into a smaller set of unique and recurring elements that they are composed of. We expect that, if successful, this approach will enable the reliable coding of a small set of basic, common "atoms" of dialogue that can both be used as standalones (similar to elementary substances), as well as to recreate a larger set of more complex, less frequently occurring categories (similar to chemical compounds).

2. Method

The current research was conducted in several stages. First, we searched for existing assessment coding tools for classroom discourse that met our criteria (stage A). Then we systematically compared the different coding tools by using all of them to analyze one lesson transcript (stage B). Following, we identified shared elements among different coding categories given to the same turns of talk, concentrating on elements that appeared in most tools (>4), and were also recognized as important in the extensive APD literature (stage C). After that, in stage D, we used these elementary particles to code other lesson transcripts and tested inter-rater reliability (stage E). Finally, we searched the coded data set for recurring combinations of specific dialogue elements and investigated how these compound combinations mapped onto the different existing coding categories from the tools reviewed (stage F). The work progress was such that stages A and B iterated until saturation was achieved. Stages C through E became a cycle that iterated several times until the emerging coding scheme was able to capture most discussion turns in the transcripts coded, while still maintaining sufficient inter-rater reliability.

2.1. Research settings, participants and data collection

The present work was situated in a larger research and development effort aiming to encourage APD in upper elementary Hebrew language arts (HLA) lessons in a large school district in Israel. The project used a collaborative research-practitioner participation (RPP hereafter) approach and harnessed expertise and knowledge of educational practitioners (8 teachers from 4 different primary schools, 5 language arts coaches) and four research team members. This development team designed 10 different lesson units based on narrative texts taken from the standardized curriculum for these age groups (for more information, see Barak & Lefstein, 2021).

The eight participating teachers had implemented lesson units in their own classrooms. A typical lesson unit would consist of four activities around a chosen text, spread out over approximately five lesson hours. These activities include a technical reading phase, a structural reading comprehension phase, an interpretative whole classroom discussion around a central open question and a writing activity.

Lessons in which the interpretative discussion took place were targeted for video and audio recording. Over the course of 1.5 school years (2018–2019, YR1 and half of 2019–2020, YR2 of the project, which was cut short by the COVID-19 pandemic restrictions), twenty whole classroom discussion protocols were chosen for analysis. Each recording involved the HLA teacher and her students (ranging from 20 to 34 students per class) and was collected by a research team member. Active parental consent forms for video and audiotaping were collected from at least 80 % of the students in each classroom. Students without active parental consent were seated outside of the scope of the camera. Thus, even though they were not excluded from the lesson activities and discussion, their answers were not recorded or transcribed.

The classroom dialogue recordings selected for the present work consisted of five different 45 min-long lessons, from four female HLA teachers (4 different schools, serving different populations), at different stages of their involvement in the project. Classroom dialogue of consenting participants was transcribed in verbatim and parsed by conversational turn. One lesson protocol was used to compare the different coding tools selected and extract basic elements (see stages B and C further on). One protocol was used for training coders, and the remaining three protocols were used for establishing inter-rater reliability (stage E further on).

2.2. Research procedure and stages

2.2.1. Stage A: selection of coding tools for review and extraction

The criteria for selecting the coding schemes for this process were as follows: They had been developed by well-known research groups or educational labs with considerable expertise on APD, for prominent studies, iterated several times, tested on large data sets of transcribed dialogue protocols obtained from K12 students or classroom recordings, used in empirical, in scale, quantitative research that focused on student learning from or through dialogue, in micro granularity.

Other selection criteria were based on the type and method of data collection and analysis, as well as the parsing strategies: We did not consider tools that were developed specifically for mathematics classes (e.g., Heyd-Metzuyanim et al., 2020), to capture dialogue outside of elementary and secondary school environments (e.g., informal education, family conversations, higher education, and kindergarten), nor to assess social-emotional aspects of classroom talk (e.g., Molinari et al., 2013). Moreover, we only included tools that used conversational turns as the unit of analysis and for which transcribed versions of classroom dialogue were the main source for coding procedures. This excluded coding tools that used larger grain sizes of dialogue, such as episodes, spells or entire discussions (e.g., Pianta et al., 2008; Reznitskaya & Wilkinson, 2017, 2021) and methods developed to code from video or audio recordings or in vivo during structured field observations.

2.2.2. Stage B: systematic comparisons of the different coding tools

To test similarities across coding schemes, we used a methodology suggested by Hennessy et al. (2016) and coded one classroom discussion transcript with all the coding tools selected in stage A. The transcript contained a total of 307 turns of speech (teachers and students combined). In this transcript the teacher and student conduct an interpretative discussion around a narrative text called “Fly, eagle, fly!” a story about a farmer who finds an eagle fledgling and raises it as a chicken, until a friend of the farmer convinces the bird itself that it is an eagle. The teacher had designed the lesson to revolve around a central question: “Why did the farmer do what he did (raise the chicken as an eagle)?”. The transcribed dialogue protocol has taken place after the technical reading phase of the narrative text. The transcript was parsed at the level of speech turns and plotted on a spreadsheet computer software, one speech turn per row and one coding tool per column. Thus, the content of each speech turn was held constant and analyzed several times (with the different selected coding tool as variables).

The coding work in this stage was conducted by the first author. The purpose was to try and unearth basic commonalities between the different existing coding categories that make up the seven coding frameworks chosen for review. By using the existing categories to code the same turns, the shared essence embedded within the differently named constructs became much clearer.

2.2.3. Stage C: extraction of shared dialogue elements across coding categories

After coding, the shared elements found were identified and extracted, using our theoretical understanding of their function or aim in the coding tool (i.e., which discourse moves, or articulation attempts they were meant to capture). The dialogue elements (DEs hereafter) we unearthed are the following: Invitation, Elaboration, Repetition, Reasoning, Challenge, Evidence, Simple, Evaluation, and Refer (see parts 3.3.1–3.3.9 for further information on each DE). In most cases, as will be shown ahead, similarities were prominent and evident even in the original categories’ names. In other cases, the prism of speech acts was useful in separating the pragmatic meaning from the semantic form of expression (Austin, 1975; Searle, 1969).

We showcase how this prism proved useful in the case of one element, Elaboration. According to the English speech act dictionary (Wierzbicka, 1987) this act is grouped under the Tell group, that is: verbs which are meant to cause someone to know something. Yet, this category has been referred to by other names as well, such as clarification (Walton, 2007), interpretation (Hancher, 1981), explanation, continuation, or expansion. Regardless of how this act is named exactly, the intent of the person performing it is to add more information to make something obscure clearer (Jasniskaja & Karagjosova, 2011). Considering differently named categories through a specific lens of speech acts, narrowed the interpretation of what the speaker meant to achieve and aided us to distill and recognize the shared, basic element that existed in the differently named categories.

2.2.4. Stage D: coding classroom transcripts with dialogue elements

To test the usefulness of coding classroom transcripts with DEs, we coded a YR1 lesson protocol. Coders were instructed to consider each DE as a standalone category and look for its existence in a conversational turn, while taking the immediate context of the turn in consideration for correct interpretation. For each of the DE categories, coders indicated ‘1’ if an element was present at least once, or ‘0’, if not. The elements represent relatively well-defined and relatively easy-to-spot aspects of the conversation.

2.2.5. *Stage E: testing interrater reliability measures*

To test the interrater reliability of the coding scheme, three different lessons from the YR2 database were selected, each from a different participating teacher and school. Two trained human judges (the first author and another researcher involved in the project) coded the total of 935 conversational turns. After coding each lesson (approx. 300 turns each), the judges met to discuss their decisions, while focusing their attention on points of contention in the coding, proceeding to code a new lesson based on improved understandings of the boundaries of each DE examined and lists of key words serving as indicators. Reliability indices were calculated on the coded transcripts, prior to discussion.

2.2.6. *Stage F: identifying recurring compound combinations*

Already during stages B and C, it became apparent that distinct combinations of identified elements make up easily recognizable "compounds" parallel to constructs of importance for APD quality assessment. For example, questions (or invitations) in general are easily recognized but are not very informative in and by themselves. The identification of one (or more) additional DE(s) in a turn enables us to differentiate between instances where a teacher invites further elaboration ("What do you mean? Can you explain that a bit further?"), from cases where she specifically invites reasoning ("Why would the farmer put the eagle in the chicken house?"), or challenges ("Is there anyone who thinks the opposite?"), and so on.

Intrigued by this simple idea, we then looked for systematic repetitions of co-occurrences of different DEs in single (already coded) turns. We then checked whether and how specific DE combinations mapped onto well-known, compound dialogue constructs from the seven existing coding tools. For example, we extracted all the instances in which coders had identified the DEs Repeat, Elaboration and Refer in a single teacher turn. In most cases, these turns captured instances where teachers repeated ideas suggested earlier by students, and further clarified or summarized them. This combination of DEs corresponds with the compound talk move *Recapping* (e.g., Felton et al., 2019).

3. Findings

3.1. *Stage A: coding tools chosen for review*

The selection procedure resulted in a set of seven coding tools: Kuhn and colleagues' coding scheme for measuring Peer to Peer Argumentation (Crowell & Kuhn, 2014; Kuhn & Zillmer, 2015; P2PA hereafter), the Low inferences Dialogue Observation tool by O'Connor and LaRusson (2014; LIDO hereafter), Scheme for Educational Dialogue Analysis (SEDA hereafter) developed at Cambridge University by Hennessy and colleagues (Hennessy et al., 2016), the Science Discourse Instrument developed by Osborne and colleagues (Fishman et al., 2017; Osborne et al., 2019; SDI hereafter), the Israeli Pedagogy coding scheme developed by Adam Lefstein and colleagues (Pollak et al., 2015; IP hereafter); the Argumentation Teacher Move coding scheme developed at Pittsburgh University (Correnti et al., 2015; ATM hereafter), and the Students Talk Move tool designed by Hardman and colleagues (Hardman, 2019a–c; STM hereafter).

Each research team generously shared detailed explanations and coder handbooks, which were helpful in operating the tools with fidelity. Table 1 lists the main characteristics of each selected scheme, based on the three dimensions detailed in Sections 1.1 to 1.3 of this paper. Further detailed information on the theoretical background, specific goals, and unique features of each of the seven tools is included in the Supplementary materials.

3.2. *Stage B: similarities and complexities uncovered*

Coding one lesson with several coding tools meant that each conversational turn was analyzed several times with different selected coding tools as variables. In some cases, more than one category (from a single tool) was found applicable (to a single turn). This procedure highlighted the shared traits across the different tools, as well as the challenges of using long and intricate coding tools, where some of the distinctions between categories were not clearly discernible. In other cases, the tools contained such a limited number of coding categories t that only very few aspects of the data set could be coded. This highlighted the data loss caused by a predetermined focus on certain participants or specific types of talk moves and forfeiting attempts to capture other aspects of the

Table 1
Overview of the seven coding tools reviewed, according to three dimensions of variance.

	No. of codable variables	Reliability test scores	What is coded?	Teacher/student/both
SEDA	33	NA	Both existing classroom talk and APD moves	Both student and teacher
ATM	29	NA	Only APD	Mostly teacher – different categories
STM	25	NA	Both existing classroom talk and APD moves	Mostly student – different categories
IP	52	NA	Both existing classroom talk and APD moves	Both student and teacher
SDI	12[a]	High	Only APD	Both – but different categories
LIDO	12	High	Both existing classroom talk and APD moves	Both – but different categories
P2PA	34[b]	High	Only APD	Only student

[a] Formally, the SDI tool features 6 coding categories, however instructions for coders and examples given in the team's papers distinguish between two stages of execution for each coding category – emerging and proficient. This then increases the number of codable variables from 6 to 12.
[b] 24 in some iterations.

discourse as well.

Table 2 showcases a small segment of the lesson coded with the seven selected coding schemes. In the first turn, the teacher asks an open ended, contestable question, with multiple potentially valid answers. The various coding tools offer differently named categories, that all code the same phenomenon. The similarities among P2PA's *Justify?* and SEDA's *Ask for justification* are present in the names themselves, but we argue that in essence the same is true for SDI's category of *Ask* (proficient execution), LIDO's category of truly open and contestable teacher question, IP's category of *Uptake* and ATM's category of *Inference and analysis question*. Since all coding tools were basically designed to capture similar things, these similarities are perhaps not surprising, yet juxtaposing the various coding categories next to each other accentuated the shared essence across coding tools.

Applying the schemes that include many coding categories often resulted in multiple possibilities from a single tool for coding a single turn. This can be seen in Table 2, in three cases indicated by the word OR, where more than one category from the same tool was applicable. The operational definitions and examples provided in the handbooks for coders were not always helpful in distinguishing among ambiguous and confusing cases and choosing the most fitting category out of the many options.

Finally, and in contrast, we would like to draw attention to the three cases of N/A in Table 2, indicating that a given coding tool did not include a relevant category for that specific turn. For example, STM focuses on student talk moves only and therefore precludes capturing teacher talk moves. Another example, of a different nature, is evident with the P2PA tool, which was designed to capture dyadic interactions between peers and therefore precludes categories for the teacher's response in turn 3.

3.3. *Stage C: dialogue elements uncovered by comparison and extraction*

After several iterations, we arrived at nine Dialogue Elements. These DEs are defined as mutually exclusive, basic "atoms" of classroom discourse features that were included in the majority of the seven coding frameworks examined (>4). Following are brief descriptions of each DE, accompanied by theoretical background and selected examples of how it appears in the coding categories from the seven coding frameworks examined, either as a discrete, standalone feature, with an additional qualifier (such as identity of speaker), or as an element of a combination with additional DEs. A complete account of how each DE appears in each of the seven coding frameworks' categories is provided in the Supplemental materials.

3.3.1. *Invitation*

Questions and invitations to participate are a common element of teacher-led classroom discourse (Coulthard, 1992; Galton et al., 1999; Howe & Abedin, 2013; TALIS 2008 Technical Report, 2010). A frequent distinction made in the teacher-led classroom dialogue literature is the difference between closed, test-like questions and authentic, open-ended questions. Teacher open-ended questions have been associated with subsequent student learning gains (Hennessy et al., 2016; Nystrand et al., 2003). Similar findings have been reported for peer tutoring settings (Chi & Roscoe, 2002). Yet, even though they are considered a distinguishing characteristic of APD (Alexander, 2018; Wells & Arauz, 2006; Wilkinson et al., 2015) they are not very common in typical classrooms (Coulthard, 1992; Galton et al., 1999; Howe & Abedin, 2013). In peer-led, collaborative group work contexts, questions have also been found to predict learning gains (Mercer et al., 1999; Rojas-Drummond & Zapata, 2004; Shachar & Sharan, 1994).

Table 2
A short dialogue segment, coded by the seven different coding tools:

	Text:	P2PA[a]	LIDO[b]	SEDA[c]	SDI[d]	IP[e]	ATM[f]	STM[g]
T	Why *did* the farmer manage to do what he managed?	Justify? Request support preceding claim evidence or argument.	T4 Teacher poses truly open, contestable question	I1 Ask for explanation or justification of another's contribution	Teacher Ask Proficient	Uptake the teacher integrate student's question in a new question addressing the classroom	Inference and analysis Q open ended **OR** Uptake/ Push back – teacher rejoinder move	**N/A**
S	Because he.. was a fledgling!	Respond The answer to a question?	S6 student's turn in simple clause or less.	R1 Explain another's contribution	Student Explain emerging	Correct students response **OR** Claim student's comments	Weak Text based evidence	S explain **OR** S justify
T	Excellent. Roni, did you hear Maor's response?	**N/A**	T1 Teacher get student to respond to another student's turn.	G1 Encourage student to student dialogue	Teacher Link emerging	Invitation to peer feedback	Invite students to link teacher rejoinder move	**N/A**

[a] Peer to Peer Argumentation coding tool, Kuhn et al.
[b] Low Inference Discourse Observation coding tool, Michaels & O'Connor.
[c] Scheme for Educational Dialogue Assessment, Hennessey et al.
[d] Science Discourse Instrument, Osborne et al.
[e] Israeli Pedagogy coding tool, Lefstein et al.
[f] Argumentation Teacher Move coding tool, Correnti et al.
[g] Student Talk Moves coding tool, Hardman et al.

All seven coding frameworks included at least one category intended to capture explicit invitations for participation, and most had several different coding categories to distinguish between different types of such requests. Inspired by the SEDA framework (Hennessy et al., 2016; Vrikki et al., 2019), we chose to name this element *Invitation*, instead of the more commonly used 'question'. The term "invitation" signals more clearly that this category also includes cases when interlocutors explicitly encourage (an)other(s) to contribute to the discussion, without using the syntactical form of a question (e.g., "I want to hear more people contributing their solutions" or "Tell us what you think"). Vice versa, participants may use the syntactical form of a question for other intents than to explicitly invite an opinion or contribution. For example, challenges to previous claims can be disguised as questions (e.g., "What? Where do you see a generous act in this text? There is none").

The DE Invitation did not appear as a standalone, discrete category in any of the seven frameworks. However, it does appear in every coding scheme with additional qualifications and/or in combination with other elements. For example, in the LIDO category T4 (teachers' open questions) the invitation is qualified by speaker (teacher only) and type (open). The SDI category ASK is defined as an open question with plural potentially valid answers. It then also refers to teachers only, but in addition requires that coders qualify the question as either proficient (an authentically open question) or emerging (closer to a test like question, with few permissible answers). The DE Invitation is often used in combination with other DEs, such as Reasoning (e.g., SEDA's Probing for reasoning) or Elaboration (SEDA's Invitation to clarify).

3.3.2. Elaboration

A key element of APD is when participants contribute more on an idea, solution or proposition, by adding more information about it, explaining it or expanding it with the intention of making sure others have understood the communication and/or the ideas it intended to convey (Howe et al., 2007, 2019; Webb et al., 2009). The aim of the speaker is not to strengthen the epistemic status of a particular claim per se, but rather to clarify a particular idea that was introduced to the conversation, whether by the original speaker or by others (Alexander, 2018; Reznitskaya, 2012).

Elaboration is the name chosen for the shared element found in differently named categories. We build on the model suggested by Asterhan and Schwarz (2009) who distinguished between turns where the new information serves to clarify the content, from more dialectical moves in which the speaker calls into question the epistemic status of a particular idea, solution or claim (see DE Challenge, in part 3.3.5 ahead) and/or from instances in which the speaker aims to justify or reason a particular standpoint, solution or idea (see DE Reasoning, in part 3.3.4 ahead). Benefits of elaborations (or: clarifications) have been documented in several research works (e.g., Chi, 1994; Osborne & Patterson, 2011; Webb, 1989). Yet, the operational definitions of this category vary extensively across coding tools. For example, some have distinguished between explanations, justifications or challenges (Alexander, 2018; Reznitskaya, 2012), while in other instances, the main distinction is between simple or brief responses and any turn that contains multiple inferences, regardless of the relation between them (Chi, 1994; Osborne & Patterson, 2011; Webb, 1989).

This trend is also reflected in the seven tools examined here. Most of them distinguish between elaboration (any information added) and justifications (specifically, added facts or reasons to support a claim), whereas others do not. For example, LIDO's category *S4* (student elaborated turn, longer than a simple clause) is applicable if the student's contribution is longer than a simple clause, regardless of intent or function. In contrast, other tools had even further distinctions, separating cases when student clarify meaning by adding more information, from cases where clarification comes from elaboration. For example: STM's categories of *SExpand/Add* and *SExplain*, and in P2PA, the categories *Add* (an extension or elaboration of the partner's preceding comment) or *Extend-O* (contribution meant to extend or elaborate the partners' preceding utterance), respectively.

In most frameworks examined, the Elaboration element appeared in tandem with a qualification to a speaker or a situation. For example, ATM's categories of *Teacher provides info & student's strong/weak explanation*. Or SEDA's categories *B1 & B2* (clarifying answers own & others), respectively. In other cases, the element was embedded within coding categories as a component in a compound construct. For example, SDI's *Teacher link*, used for coding turns where the teacher adds more information, facts or outside knowledge to help with the tasks, this speech act must include an element of elaboration. Finally, the element of Elaboration also appears as a standalone, for example in IP's categories of *procedural* or *topical* explanations.

3.3.3. Repetition

The element Repetition captures instances in which a speaker reiterates something that was already mentioned during the discussion. The repetition may be of an idea previously introduced by him/herself or by someone else, and it can be nearly verbatim or in a rephrased form. Mere repetitions would not be considered a defining characteristic of APD, and too many simple repetitions could even indicate stagnation (Bleicher et al., 2003). Yet, this aspect of the discourse is still coded due to its centrality in typical classroom talk. For example, a teacher may repeat task instructions to make sure everyone caught on or relaunch a question to invite more (different) responses. A student may repeat a contribution if they believe it was not heard or considered.

When using the lens offered by the elements to compounds approach, Repetitions can also be seen as a component in many APD types of talk moves. For example, when a teacher recaps a range of different ideas that were proposed up until a certain point in the discussion, he or she is, among other things, is repeating what was already said. The same is true for a participant reformulating ideas, recontextualizing, reframing, reaffirming or revoicing.

Some coding tools featured the element of Repetition as a standalone, for example, in STM the category of *Srep* (for student rephrasing, repeating, or reformulating), in IP's category *Repeat* and in ATM *Repeat*. In other coding tools the element was embedded within compound combinations, for example, ATM's or IP's *Recap*, meant to capture instances when a teacher stops, links together and clarifies several ideas and contributions previously mentioned. Or ATM's *Reinitiate*, which is used to capture instances where the teacher repeats the main question in attempt to get more, or different responses (also known as relaunch). In SEDA, the element of

Repetition appears as a component in combinations with other elements, for example, *P1 synthesize ideas*, *C1 refer back* and *C2 make learning trajectory explicit*, or even *B2 build on other's contributions* necessitate repetition at least of some parts of what was said in preceding turns.

3.3.4. Reasoning

Most dialogic learning theories link the academic productivity of classroom discourse to the quantity and quality of reasoning evident in both the teacher and the student contributions (e.g., Howe et al., 2019; O'Connor & Michaels, 1993; Resnick et al., 2018b; Reznitskaya et al., 2009a, b). A body of empirical evidence supports this notion and associations between improved individual reasoning and learning gains has been found (Chinn et al., 2000; Wegerif et al., 2004).

In the APD literature, Reasoning refers to a social practice, rather than the set of formal rules of logic (Keefer et al., 2000). It enables students to externalize their own thought process and internalize other ideas and different perspectives in an open and free encounter (Mercer et al., 1999). The element Reasoning captures instances in which the speaker explicitly expresses an opinion, shares the reasons of his thinking, or explicates the inner thinking process that brought him/her to a conclusion.

The DE Reasoning was found in each of the seven coding tools examined. As is the case for the element Invitation, it rarely appears as a standalone, discrete coding category, but rather in combination with other elements (such as Invitation, Referring or Evidence), or with further qualifications. For example, in LIDO's category of *S3* (student's use of reasoning to support their claim) and in IP's categories of *Student claims* and *Teacher claims* the reasoning is qualified by the speaker (a student or a teacher). Reasoning is also often combined with other elements too. For example, in SDI's *Student critique* the element of Reasoning is not only qualified by a speaker (student), but also combined with Refer (see Section 3.3.9) and Challenge (see Section 3.3.5), respectively. Other examples include STM's *SJUS* (student's justification: Reason + Evidence), or ATM's categories *Pressing for reasoning* (Invitation + Reason) and *Teacher thinking out loud* (Reason + Elaboration).

3.3.5. Challenge

This element captures instances when an interlocutor explicitly challenges the epistemic status of a particular idea, solution or claim that had already been contributed to the discussion space. This can take many forms, ranging from subtle challenges to more direct and articulated oppositions, counterarguments, and rebuttals (Asterhan & Schwarz, 2009; Wilkinson et al., 2015). These turns can include personal references to the individual(s) who voiced the challenged proposition ("I have a different opinion from Sylvia's") or phrased in more neutral terms ("The question is, however, how do we really know that the earth is round?"). They can be reasoned (with elaborate counterarguments or rebuttals), or not ("I do not agree"). What these all have in common is that in a Challenge, we capture the speaker's opposing stance toward a previously contributed idea, whether he/she explicitly doubts, questions, queries or disagrees with it.

Different models of APD vary in their emphasis on the critical dimensions of classroom discourse. Specifically, those originating from argumentation and/or socio-cognitive theories tend to attribute it a more pivotal role and often value detailed distinctions between different types of Challenges (authors, 2009; Crowell & Kuhn, 2014; Nussbaum, 2011). Yet, regardless of differences in emphases and detail, the active exploration of different views, solutions, interpretations or idea is a central feature of all prominent models of classroom APD (Wilkinson et al., 2015). This is evident, for example, in Mercer et al.' (1999) distinction between cumulative and exploratory talk, in classroom discussion-based reading comprehension approaches (Murphy et al., 2009), and in discussion-based mathematics and science lessons where teachers encourage students to generate and compare different solutions (Larrain et al., 2019; Webb et al., 2009, 2014).

In all but one of the seven frameworks examined, there was a discrete standalone category for this element, such as in SEDA's *P5 challenge viewpoint*, STM'S *Schal*, PSPA's *Disagree* and IP's *Explicit disagreement*. We also found instances of Challenge with additional qualifications, such as IP's *Challenge to the teacher* which is confined to specific situations in which a student challenges a teacher's statement. Challenge was also embedded within categories as a component combined with another element: For example, SEDA's *I2 invite disagreement* is a combination of the elements Invitation and Challenge. ATM's category of *Uptake/push back* is meant to capture instances where the teacher either picks up the idea and builds on it (uptake), *or* critically examines it and returns it to the speaker for further scrutiny (push-back). In the element to compound combination approach, uptake can be expressed by a combination of Refer and Elaboration, whereas "push-back" can be recreated by combining Refer and Challenge.

3.3.6. Evidence

The next element, Evidence, captures instances where the speaker uses facts, specific quotes from the text, or external knowledge sources to support his/her claim. Unlike the previously discussed element of Reasoning, which captures instances of inference, here the shared element under inspection is the speaker's use of *external knowledge sources*, such as quotes or facts based on experts, text or data.

Several theoretical and conceptual models consider use of evidence as one of the defining features of productive talk (Kuhn & Udell, 2003; Resnick et al., 2018a). It is a key element of productive argumentation (Rapanta, 2019), as students learn to evaluate the strength of evidence and distinguish between stronger and weaker support for their arguments (Osborne, 2010; Osborne & Patterson, 2011; Reznitskaya et al., 2009a, b).

In most frameworks use of evidence is coded as a standalone category, such as in, for example, STM's *Sjus* (students justifying their claim by providing reasoning or evidence) or ATM's *Strong text-based evidence*. In some cases, the use of evidence is qualified to a person, for example, in LIDO's *S3* (students providing evidence), or in SEDA's categories of *R1 & R2* (provide evidence to own or other's claim) which differentiate cases of own or others. In other cases, the use of evidence was embedded within the category. For example, SDI's category of *Press* (encourage students to support their answer with reasons and evidence) is a combination of the

DECCA elements Invitation, Reasoning and Evidence. This is also true for SEDA's category of *I4* (ask for justification) and P2PA's *Justify?* (a request of the partner to support their claim with evidence). In the IP framework, the category of *Teacher claim*, is meant to capture instances where the teacher weighs in and changes the course of discussion using authority, knowledge, and facts. Therefore, there is a strong sense of the Evidence element, although it is not present in the category's name.

3.3.7. Simple

The element Simple is meant to capture a prominent aspect of students turns in everyday classroom talk, namely very short reactions, comments, or answers. These expressions are presented by the 'R' (response) in the IRE or IRF model of classroom discourse (Coulthard, 1992), and are commonly typified as answers to "closed" teacher questions (Howe & Abedin, 2013; Peled-Elhanan & Blum-Kulka, 2006). As its name suggests, this category does not represent necessarily APD type of responses and is therefore not discussed much in the APD literature, other than to notice its prevalence in everyday classroom talk and contrasting it with APD. Still, most of the tools reviewed here include one or more categories to capture this element. The only tool that specifically chose to ignore non-APD talk moves is SDI, but in their instructions of coders they distinguish between cases of *proficient* and *emerging* turn styles. Therefore, the element of Simple exists, but in a slightly different capacity.

Some coding tools distinguish Simple as a discrete coding category, LIDO's category of *S6* (turn is simple clause) or STM's category of *BSC* (brief student contribution). In other tools, the element is not present in the categories' names, but nonetheless, the instructions for coders indicate that it is meant to capture relevant, but less sophisticated answers. For example, in PSPA's categories of *respond* (the answer to a simple question, which does not advance or clarify the speaker's position).

As an elementary particle, Simple is inherently different from the other elements extracted. It is neither a recognized construct of importance in APD, nor a speech act or a talk move. In a way, it even stands as a contradiction to the rest of the elements and their combinations, because once a turn is categorized as simple, it cannot contain more than one element (such as Reasoning, Evidence or Elaboration). That being said, one idea that we have tested rigorously was using the Simple element as a qualifier, meaning that attaching it to other elements will indicate the turn as a less proficient or even a failed execution of the intended elementary particle. Such as, for example, wrong use of reasoning, misuse of textual evidence, failed elaboration attempts and so on. However, we were unable to reach sufficient inter-rater agreement indices on this specific use. We suspect that execution of this function necessitates much higher inference than is required by the rest of the elements in our tool kit, and was, therefore, more confusing for coders.

In fact, the only combination found that maintained the low inference aspect of the other elements suggested, is the combination of *Invitation* and *Simple* to represent closed, test-like questions that have a limited range of clearly defined correct answers known to the teacher in advance. Failing to use the *Simple* element in combination with other elements caused us to ponder whether it should even be included in this list. However, as we have argued before, if one is interested in capturing a process of change, from traditional classroom discourse to more APD type of discourse, the coding tool must be able to capture both states with fidelity. Simple, short, factual answers and comments are such a dominant aspect of everyday classroom discourse that we felt it should not be ignored.

3.3.8. Evaluation

The element Evaluation captures instances of direct feedback or passing a judgment on something that was already said (Berry & Kim, 2008). In a classroom context, it is typically the teacher who is in a position to evaluate the correctness, relevance, or validity of her students' answers. There is a long tradition in some branches of APD of coding different types of evaluations with specifications of whether it was positive or negative, implicit or explicit and even correct or incorrect.

Yet, when trying to implement more academically productive discourse, these evaluations of student turns can be complex. On the one hand, successful APD implementations are measured, among other things, by an increase in number of active participants. When teachers provide constant evaluative feedback of student contributions, it is likely to suppress participation by students with different viewpoints and/or with less academic self-efficacy (Segal & Lefstein, 2016), as well as unintentionally signal students that only those with the 'right answer' can participate (Clarke et al., 2016). On the other hand, critically challenging evidence and claims is an important aspect of argumentation and reasoning (Chiu, 2004). We then follow distinctions proposed by Ford and Forman (2006) and capture teachers' active navigation and control of the discourse *form* with the element Evaluation. Teacher actions of assuming or modeling the role of *constructor* are captured with the element Elaboration, on the other hand, and the role of *critique* is captured with the element Challenge.

Some coding schemes distinguish Evaluation as a discrete coding category. Examples are STM's category *sEval* (students evaluate and make judgements), ATM's category *Terminal move* (in which teachers end the exchange by commenting on its quality), or in SEDA's *G4* (providing informative feedback). In other cases, the element of Evaluation was embedded within combinations with other elements, such as Invitation in SEDA's *I2* (invite evaluation) or Refer (see Section 3.3.9) in SEDA's *P2* (evaluate alternative views). Next, and despite its name, P2PA's *Coopt* is intended to capture acknowledgements, explicit or implicit, that the partner's preceding utterance is correct or supported. This then also includes a clear element of evaluation. Finally, in IP, there are numerous and several different types of evaluations defined by execution style (explicit or implicit), content (negative, positive, neutral and combined), as well as qualified by speaker (teacher or student).

3.3.9. Refer

A final important feature of APD found is when learners engage with and interact on the contributions, ideas, and ways of thinking that others add to the conversation (Webb et al., 2014). In the Accountable Talk framework, for example, this is expressed in the notion of accountability to the learning community (Michaels et al., 2008; Resnick et al., 2018b). In Alexander's approach to dialogic teaching (2018) this element appears in four of the five principles of dialogic teaching (i.e., collective, reciprocal, supportive and cumulative).

Similarly, the exploratory talk framework (Mercer et al., 1999; Mercer & Littleton, 2007), posits that student learning improves when they listen and respond to others and when they build on the ideas of fellow learners.

We have named this dimension of the dialogue as Refer. The element is meant to capture instances in which a speaker explicitly addresses, relates, links, or connects ideas, contributions or opinions that were previously uttered to the content of the current contribution (i.e., "I think Roni's solution is the better one"). In this case, the speaker does not repeat what Roni says, as the classroom already heard it, and while addressing Roni, the speaker is clearly referring to the context of her earlier claim, not her per se. These activities share the fact that the speaker explicitly acknowledges and relates to (a) previous contribution(s) in the dialogue.

In some coding frameworks, this element is coded as a standalone category, as is evident in the SDI category Link or SEDA's *C1* (refer back). In other frameworks, the concept is conditioned by a speaker, quality, or a situation. For example, ATM's *Strong/Weak Link* aims to capture instances where students connect their contributions and show how they relate to one another. Other examples include STM's *SCon* (students connects, refers to something else), and IP's *Direct* (student responds directly to another). The LIDO tool separates between students explicitly mentioning a person (*S1* students addressing one another) and turns where the speaker refers to the content of a previous contribution (*S2* student refers to another contribution). The only framework that did not include the DE Referring is P2PA. However, this framework was specifically developed for coding argumentative discussions in dyads or very small peer groups, instead of whole classroom settings. This has likely rendered this interactive-transactive feature redundant.

3.4. Stage D: coding classroom talk with DECCA

Table 3 shows how DECCA was used to code classroom dialogue transcripts. Space limitations prevent us from sharing more and longer examples, but we hope that the selected excerpt is rich enough to showcase both the use of DEs as standalones, as well as in combinations.

The first three turns, contributed by the teacher and two participating male students, are pretty straightforward and can be captured with the elements used as standalone. In the first turn, for example, the student provides a reason for a story character's behavior, as is therefore coded with the DE Reasoning. In response, the teacher further elaborates and explains features of the student's answer, and this turn is therefore coded with the DE Elaboration. In the next turns, the teacher, as well as the student who answer her, are performing more complex moves. For example, the teacher not only invites a random student to state his reasoning, but she also specifically searches for S2 and invites him to refer to the first speaker. Thus, DECCA coders will indicate that this turn includes the elements Invitation, Reasoning and Refer. If coders would have focused on compound, higher-order dialogue move constructs, they may have identified it as 'probing for reasoning', 'invitation for alternative opinions', or 'opening the dialogic space'. Subsequently, S2 then not only challenges the previous student's contribution, but also provides reasoning to support his own claim, thereby making a rare, yet important move often called 'reasoned counter-argumentation'. Here too, instead of searching for such rare and high-inference coding category that might only be applicable in far-between cases, DECCA coders only search for and code the basic dialogue elements in a turn (i.e., Challenge and Reasoning in this case). When considering the two aforementioned hurdles for achieving sufficient interrater reliability (i.e., extensive coding expertise and rarity of specific talk moves), this example shows how DECCA circumvents both issues.

3.5. Stage E: inter-rater agreement

Table 4 shows Cohen's Kappa interrater reliability indices, the raw percentage of interrater agreement and the number of times an element appeared in the data set (only cases that both judges agreed on were counted). Within the limitations of the data available, we tried to increase the variety of talk moves as much as possible. Firstly, by picking lessons from three teachers from different schools, and second, we chose only lessons from YR2, assuming that by then more APD type of talk moves will be integrated into the discourse, allowing us to capture more execution forms of it.

In comparison, the DEs that are specific to APD talk moves (Reasoning, use of Evidence, Refer and Challenge) were less frequent in the data set, but still had good-to-satisfactory Cohen's κ and high agreement rates. The one exception is the DE Challenge for which, despite high raw agreement rates (96 %), the Cohen's κ interrater reliability proved only acceptable (.65). This element was also one of

Table 3
An example of how DECCA was used to code whole classroom dialogue.

Speaker	Text	Invite	Elaborate	Reason	Refer	Challenge	Simple	Repeat	Evidence	Evaluate
S1	Ermmm… like.. it… he doesn't know how to cry because he didn't cry in all of his life. It's something that-/you have to get used to that.	0	0	1	0	0	0	0	0	0
Teacher	That means something that you would have to practice.	0	1	0	0	0	0	0	0	0
S2	No!	0	0	0	0	0	1	0	0	0
Teacher	Who said that? S2? Can you tell me why not?	1	0	1	1	0	0	0	0	0
S2	Because you don't need to know how to cry, it happens, just by itself.	0	0	1	0	1	0	0	0	0
Teacher	That means: automatically.	0	1	0	0	0	0	0	0	0

Table 4
Reliability measures of the nine DECCA dialogue elements.

Element	Cohen's κ	% of agreement	Raw frequency in data set[a] (N = 935 turns)
Invitation	0.85	95 %	171
Elaboration	0.70	87 %	226
Reasoning	0.70	97 %	34
Refer	0.74	97 %	41
Simple	0.70	90 %	150
Evaluation	0.81	98 %	33
Repeat	0.75	93 %	131
Evidence	0.77	96 %	71
Challenge	0.65	96 %	36

[a] Only cases agreed on by both judges were counted.

the least frequent in the data set ($N = 36$, or in 3.8 % of the speaker turns), which could in part be responsible for the lower reliability. Yet, it cannot be the whole story, as the DEs *Evaluation* and *Reasoning* were even less frequent, but produced better interrater reliability. Another explanation could be that in whole classroom discussion, acts of challenge are more difficult to identify, as they are often 'hidden' in various forms and shapes of verbal communication, to soften potential social ramifications and to avoid face threat (Asterhan, 2013; Asterhan & Babichenko, 2015; Brown & Levinson, 1987; Geoffrey, 1983).

It is also possible that challenges are less easily identified in different cultural settings. Vrikki et al. (2019) recently reported higher reliability for the Challenge cluster in SEDA-CON when coding British classroom dialogue, compared to when it was applied to dialogue in Mexican classroom. Regardless, the high raw agreement rates reported here leave room for optimism that interrater reliability on this element can be improved with more practice and more data points.

3.6. Stage F: dialogue elements combinations reconstruct APD coding categories

In the final stage, we mapped how these nine DEs can be used, not only as standalones, but also in combinations to capture additional classroom dialogue features and reconstruct many coding categories taken from the seven existing coding schemes. As an example of how this works, and for the sake of simplicity, we consider here just three elements: Invitation, Reasoning and Evidence (see Table 5). On the actual protocols of classroom discourse data, each speaker turn appears as a line, and the different standalone DE categories are juxtaposed in columns. For each turn, an element is coded as 1 when it is present, or 0, if not. In Table 5, the rows detail the APD constructs and top-down coding categories, obtained from the seven coding tools included. The columns represent the different DEs. The example shows how combinations of these three DEs can reconstruct six different classroom discourse constructs that can be used to capture both teacher, as well as student talk.

For example, we can distinguish between three different types of (student or teacher) contributions: Reasoning is coded when students or teachers expose their thinking process, by sharing how they arrived at a certain conclusion, engage with other points of view (i.e., compare or contrast) or provide reasons or causes for a proposition. Other types of contributions may only include a piece of evidence or an external knowledge source that aligns with it. Finally, when participants provide justifications, it can be captured with the indication of both Evidence and Reasoning in a turn. In combination with the DE Invitation, three different types of (student or teacher) questions can be identified: Invitation for reasoning, invitation for evidence and invitation for justification. As shown in Table 5 in further detail, most coding tools include discrete coding categories for these question types.

This example included only 3 of the DEs and serves to describe how such 'particles' of distinct dialogic nature may be grouped into meaningful compound combinations representing more diverse and complex dialogue moves. With the full set of nine DEs identified here, we were able to reconstruct dozens of distinguished dialogue constructs and the majority of the coding categories taken from the seven tools reviewed (see the second table in the supplemental materials for a complete list of all combinations found). Among these are compound combinations representing over twenty different types of questions, ranging from test-like or procedural questions to

Table 5
Using three dialogue elements to recreate compound constructs through combination coding.

Coding categories from the seven existing coding tools	Invitation	Reasoning	Evidence
Invite reasoning: Probe? (P2PA); Teacher Open Q (LIDO); Invitation for Reasoning; Reasoning (SEDA); Teacher press emerging (SDI); Press, Uptake (ATM); Open student question (STM); Open teacher Question, Essential student question (IP)	●	●	
Invite evidence: Justify? (P2PA); Invitation 4 (SEDA); Teacher Ask emerging (SDI); Open form literal function Question (ATM)	●		●
Reasoning: Counter A (P2PA); Student elaborate answer (LIDO); State position (SEDA); Student Speculate (STM); Think aloud (ATM);		●	
Evidence: Connect 3 (SEDA); Student Strong/weak text-based evidence; Teacher Link (SDI); Student linking to outside knowledge.			●
Justification: Justify (P2PA); S3 (LIDO); Student justify (STM); Student critique emerging (SDI).		●	●
Invite justification:; Teacher Press - proficient (SDI); Inference and analysis question (ATM); Probing (IP)	●	●	●

open and authentic questions. It is possible to reconstruct combinations to help capture different repetition moves, from simple repetitions or articulation corrections to explaining, expanding, paraphrasing, recapping, recontextualizing or re-voicing. It is important to emphasize that all this is possible, in addition to using the DEs as standalones, as well as capturing aspects of existing and common talk moves in the classroom, such as simple short answers and evaluations. Moreover, as coding is done at the DE element level, inter-reliability indices will not be affected by this richness and multitude of higher-order categories.

Fig. 1 plots the number of coding categories that could be recreated with the Dialogue Elements to Compound Constructs Approach (DECCA hereafter), for each of the seven coding frameworks included. The majority of coding categories could be re-created, ranging from 55 % (IP) to near-complete reconstruction rates of 92 % (SDI). The relatively lower percentages for coding tools such as IP, SEDA and P2PA is explained by the fact that these tools focus on additional dimensions of dialogue that were unique to them and not shared by the other frameworks. IP, for example, contains many categories that capture social-emotional aspects of classroom discourse. Both P2PA and SEDA contain coding categories that include elements of meta-talk (talk about talk). As we focused on the commonalities across at least four existing frameworks, we did not include DEs to capture unique classroom dialogue features. However, should scholarly interest in these (or other) dimensions increase, new elements could be added (such as, *Meta-talk*).

4. Discussion

Research on academically productive classroom dialogue is burgeoning. Yet, despite a broad conceptual consensus about the object of investigation (APD), this is not mirrored in the various ways in which it is operationalized and measured in quantitative research efforts. This lack of operational consensus poses considerable challenges for communicating, comparing and integrating research findings and, ultimately, stymies theory development and refinement.

In the present study, we joined recent attempts to develop a more unified approach for coding classroom dialogue that builds on the existing knowledge base that has accumulated thus far, and would be suitable for exploring a variety of research questions and for a range of different approaches. We also posited that, for such an endeavor to succeed, an assessment tool should be comprehensive and sensitive to detailed differences in classroom discourse features, yet reliable and consistent in its measurement, as well as capture both student and teacher talk moves, and both APD and more traditional classroom talk features with one single set of variables. We showed that in existing coding approaches these requirements are in tension and are therefore only partially met. A new and slightly different coding approach was then developed that could meet these requirements within one single coding tool: Dialogue Elements to Compound Constructs Approach.

We selected seven well-established and validated classroom dialogue coding frameworks and, through an iterative process of comparison, deconstruction and application, identified a set of basic, elementary particles of classroom dialogue, which we called Dialogue Elements. These elements are common across the different coding categories within these frameworks: Invitation, Elaboration, Repetition, Reasoning, Challenge, Evidence, Simple, Evaluation and Refer. We further argue that by coding this small set of elementary, yet frequently occurring particles of classroom dialogue moves, the number of coding categories could be held to a minimum and inter-rater reliability thresholds met. At the same time, a much larger set of "compound-like" dialogue constructs can be recreated post-coding, by identifying occurrences of specific elements appearing together at the same turn.

This small change in approach solves a hitherto common conundrum in dialogue coding development efforts: In prior work, limiting the number of coding categories for the sake of stable measurement (reliability) was often done at the expense of sensitivity and scope of the dialogue features that could be captured (and vice versa). DECCA enables the reliable coding of a small set of basic,

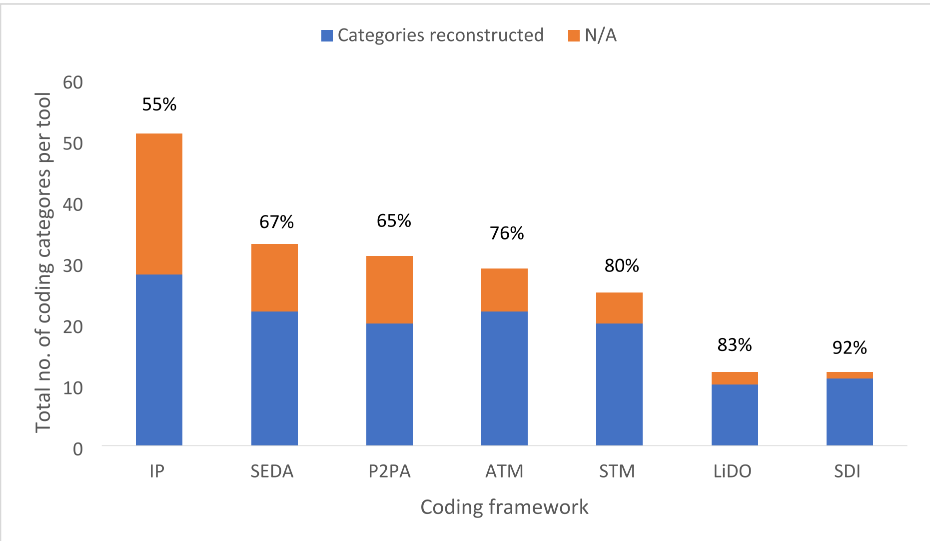

Fig. 1. Percentage of coding categories reconstructed with DECCA, for each of the seven coding frameworks included.

commonly occurring “atoms” of dialogue that can both be used as standalones (elementary substances), as well to recreate a broad range of compound, and often times less frequent, dialogue moves, such as, for example, teacher revoicing or student reasoned counter-argumentation. This change in approach also enabled the use of one set of categories for capturing both student, as well as teacher talk moves. Finally, since the more sophisticated APD talk moves are in fact made up of different combinations of a small set of elementary particles, it can also be used to capture classroom dialogue in different stages of development (from traditional recitation to “advanced” APD) with the same small set of elementary coding categories.

4.1. Previous attempts to develop a unified coding method

We were not the first to notice the plethora of different classroom dialogue coding tools, nor the first to labor toward a unified coding approach. Our efforts were heavily influenced by and built on the progress made at Cambridge University’s CEDIR group, who developed the Scheme for Educational Dialogue Assessment (SEDA; Hennessy et al., 2016). The development process of SEDA continued for several years and entailed, among other efforts, an extensive and inclusive review of the different coding approaches available, the construction of a substantial database of classroom dialogue recordings from a large number of different schools, and evaluations of the coding manual by experts from different research traditions. This ultimately resulted in a set of 33 coding categories that can capture a large palette of different dialogue moves, are applicable to teachers and students alike, and can capture both APD, as well as more mundane classroom dialogue moves. Out of the aforementioned six demands, five were then successfully met in SEDA. However, this came at a price: Interrater reliability thresholds could not be met for many of its categories.

Consequently, when SEDA was used for large-scale, quantitative research, a condensed version of the original SEDA tool was developed (SEDA CON; Hennessy, 2020; Vrikki et al., 2019). In SEDA CON, inter-rater reliability issues were solved by grouping the original 33 classroom dialogue features into eight new clusters, for which inter-rater reliability thresholds were met. This solution is satisfactory from a statistical point of view. However, it came at the expense of the tool’s sensitivity and comprehensiveness, as it now distinguishes only between eight coarse dialogue features. For example, instead of separate categories for turns in which a speaker i*sevaluating alternative views*, *proposing resolutions* or *challenging a viewpoint*, these (along with several other) dialogue moves are lumped together in SEDA CON’s cluster category of *positioning and coordinating* without further distinction.

There are many resemblances between DECCA and SEDA CON: They were developed with similar premises and aims. They both build on the considerable body of existing work on coding classroom dialogue. They both contain a small set of reliable coding categories that capture the basic features of dialogue. Even some of their respective categories’ names show a strong resemblance (e.g., SEDA CON’s *Make reasoning explicit* and DECCA’s *Reasoning,* SEDA CON’s *Connect* and DECCA’s *Refer*). The main difference is, however, that DECCA offers a way to avoid sacrificing sensitivity and comprehensiveness at the altar of interrater reliability. The nine DEs can be used as standalones to assess common, basic elements of classroom discourse, similar to the SEDA-CON’s clusters. Yet, in addition, they can also be used to recreate a wide variety of more specific and less frequent classroom dialogue features, including most of the 33 original SEDA coding categories (see the Supplemental Materials for a full list of all reconstructed categories). This, in turn, opens up new ways to empirically explore a variety of research questions that could hitherto not be explored in quantitative classroom dialogue research.

4.2. Limitations and future directions

One limitation of the current work is that we based our development efforts on a specific selection of seven existing coding frameworks. It is therefore possible that some elements of classroom dialogue were left out of DECCA. As the number of different coding tools available is nearly as large as the number of empirical publications that have applied them, limiting the number of coding tools was unavoidable. We justify our choice by purposefully selecting well-developed coding frameworks from experienced APD research groups and efforts that had been used in multiple quantitative publications and were characterized by high content validity (i. e., aligned with current theoretical descriptions of APD features). We argue that it is therefore not very likely that a different set of initial coding frameworks would have resulted in a different set of basic DEs. However, this possibility cannot be ruled and should be tested in future research.

Secondly, the DEs that were identified in are based on our understanding of the emerging consensus in the APD assessment literature. If a particular element was deemed important enough to be included in the majority of the seven coding tools, we assumed that it represents a key feature of the APD constructs and was thus included. Whereas we consider this as a major strength of the current work, one could also argue that in order to break new ground in classroom dialogue research, additional features should be included that have hitherto been considered by few only. Dimensions of classroom dialogue that cannot be captured with the current set of nine elements are, for example, rhetorical, affective and motivational aspects of classroom interaction. Future research could explore the development of reliable and valid categories for such features, if deemed of interest to APD scholars.

A third limitation is our focus on a particular level of granularity for coding, namely the conversational turn. This grain size seems to be the most common choice for quantitative APD analysis approaches. It is convenient for coding multi-participant classroom dialogue, as segmentation is provided by the participants themselves and does not have to be superimposed by coders and their subjective interpretations and decisions. However, some dialogue qualities may not be captured at this grain size. We then acknowledge the potential benefits of complementing turn level analyses with those of a larger grain size, such as conversational episodes or entire discussion activities (e.g., Howe et al., 2019; Pianta et al., 2008; Reznitskaya & Wilkinson, 2017, 2021).

Fourth, the methodology suggested in DECCA, as well as its current nine dialogue elements, should be tested in additional cultures and settings. Six out of the seven existing coding approaches we included were developed and tested on data obtained from English-speaking populations. DECCA was developed on data collected in Hebrew speaking, Israeli classrooms. Cultural differences in classroom discourse are expected (Kunh et al., 2011; Weinberger et al., 2013), and the universal applicability of any coding framework is likely to be compromised by such differences. We would argue, however, that a coding approach that focuses on the elementary, basic particles of discourse (such as DECCA) is likely to be least affected by cultural and situational differences in classroom discourse, in comparison to to coding approaches that search for compund, higher-order constructs. These expectations should be empirically tested in future research focusing on cross-cultural performance of different coding approaches (by DECCA and others).

Finally, the focus of the current work is methodological, and therefore offers little insights for educational practitioners at this stage. However, theoretical and empirical progress in any research field is contingent on our ability to accurately, reliably and efficiently capture the phenomena studied, in ways that can be communicated to, shared with and replicated by the rest of the research community. Thus, we argue that better methodology leads to better research and theory, which, ultimately, will be able to inform practice more accurately.

4.3. Conclusion

Coding classroom dialogue is a resource-intensive task and subject to human interpretation. The coding approach we have presented here (DECCA) does not alleviate that. It does, however, propose a common currency for empirical scholarly work on academically productive classroom dialogue, which is a necessary requirement for integrating findings from this very productive, yet fragmented field of research and for making further progress on classroom dialogue theory and research. DECCA is capable of capturing a wide variety of dialogue features (comprehensive), in a stable and consistent way across settings and raters (reliable) and is suitable to serve a large variety of research questions and analyses (flexible). The development of this coding approach was made possible due to the accumulated knowledge, conceptual maturation and methodological progress made by scholars of classroom dialogue in the last two decades. We do not consider DECCA to be an end destination, or as the ultimate way of coding classroom dialogue. Rather, we hope that both the questions and challenges that we have formulated, as well as the potential solution we have proposed, will further the discussion and progress in a collective effort toward a more unified approach to quantitative research on dialogic learning and teaching.

Financial acknowledgement

The research presented here was financially supported by the Israel Science Foundation (grant # 2693/17).

Data availability

No data was used for the research described in the article.

Acknowledgments

We thank the practitioners and researchers who participated in the project without whom the invaluable classroom dialogue data for this study could not have been collected. The research presented here was financially supported by the Israel Science Foundation (grant # 2693/17).

Supplementary Data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.lcsi.2023.100717.

References

Alexander, R. (2018). Developing dialogic teaching: Genesis, process, trial. *Research Papers in Education, 33*(5), 561–598. https://doi.org/10.1080/02671522.2018.1481140

Asterhan, C. S. C. (2013). Epistemic and interpersonal dimensions of peer argumentation: Conceptualization and quantitative assessment. In M. Baker, J. Andriessen, & S. Jarvela (Eds.), *Affective learning together social and emotional dimensions of collaborative learning* (pp. 251–271). New York: Routledge, Taylor & Francis Group.

Asterhan, C. S. C., & Babichenko, M. (2015). The social dimension of learning through argumentation: Effects of human presence and discourse style. *Journal of Educational Psychology, 107*(3), 740–755.

Asterhan, C. S. C., & Schwarz, B. B. (2007). The effects of monological and dialogical argumentation on concept learning in evolutionary theory. *Journal of Educational Psychology, 99*, 626–639. https://doi.org/10.1037/0022-0663.99.3.626

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

Asterhan, C. S. C., & Schwarz, B. B. (2016). Argumentation for learning: Well-trodden paths and unexplored territories. *Educational Psychologist, 51*(2), 164–187.

Austin, J. L. (1975). *How to do things with words*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198245537.001.0001

Barak, M., & Lefstein, A. (2021). Opening texts for discussion: Developing dialogic reading stances. *Reading Research Quarterly*, 1–20. https://doi.org/10.1002/rrq.413

Berry, R. A. W., & Kim, N. (2008). Exploring teacher talk during mathematics instruction in an inclusion classroom. *The Journal of Educational Research, 101*(6), 363–378. https://doi.org/10.3200/JOER.101.6.363-378

Bleicher, R. E., Tobin, K. G., & McRobbie, C. J. (2003). Opportunities to talk science in a high school chemistry classroom. *Research in Science Education, 33*(3), 319–339. https://doi.org/10.1023/A:1025480311414.

Boyd, M. P., & Markarian, W. C. (2015). Dialogic teaching and dialogic stance: Moving beyond interactional form. *Research in the Teaching of English, 49*(3), 272–296. https://doi.org/10.1080/09500782.2011.597861

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

Chi, M. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*(3), 439–477. https://doi.org/10.1016/0364-0213(94)90016-7

Chi, M. T. H., & Menekse, M. (2015). Dialogue patterns that promote learning. In L. B. Resnick, C. S. C. Asterhan, & S. N. Clarke (Eds.), *Socializing intelligence through academic talk and dialogue* (pp. 263–274). Washington, DC: AREA.

Chi, M. T. H., & Roscoe, R. D. (2002). The processes and challenges of conceptual change. In M. Limon, & L. Mason (Eds.), *Reconsidering conceptual change: Issues in theory and practice* (pp. 3–27). Dordrecht: Springer. https://doi.org/10.1007/0-306-47637-1_1.

Chinn, C. A., O'donnell, A. M., & Jinks, T. S. (2000). The structure of discourse in collaborative learning. *The Journal of Experimental Education, 69*(1), 77–97. https://doi.org/10.1080/00220970009600650

Chiu, M. M. (2004). Adapting teacher interventions to student needs during cooperative learning: How to improve student problem solving and time on-task. *American Educational Research Journal, 41*(2), 365–399. https://doi.org/10.3102/0002831 2041002365

Chiu, M. M. (2015). Statistical discourse analysis. In L. B. Resnick, C. S. C. Asterhan, & S. N. Clarke (Eds.), *Socializing intelligence through academic talk and dialogue* (pp. 301–314). Washington, DC: AREA. https://doi.org/10.3102/978-0-935302-43-1_24.

Clarke, S. N., Howley, I., Resnick, L., & Penstein Rosé, C. (2016). Student agency to participate in dialogic science discussions. *Learning, Culture and Social Interaction, 10*, 27–39. https://doi.org/10.1016/j.lcsi.2016.01.002

Correnti, R., Matsumura, L. C., Walsh, M., Zook-Howell, D., Bickel, D. D., & Yu, B. (2021). Effects of online content-focused coaching on discussion quality and reading achievement: Building theory for how coaching develops teachers'adaptive expertise. *Reading Research Quarterly, 56*(3), 519–558. https://doi.org/10.1002/rrq.317

Correnti, R., Stein, M. K., Smith, M. S., Scherrer, J., McKeown, M. G., Greeno, J. G., & Ashley, K. (2015). Improving teaching at scale: Design for the scientific measurement and learning of discourse practice. In L. B. Resnick, C. S. C. Asterhan, & S. N. Clarke (Eds.), *Socializing intelligence through academic talk and dialogue* (pp. 315–332). Washington, DC: AREA. https://doi.org/10.3102/978-0-935302-43-1_25.

Coulthard, M. (1992). *Advances in spoken discourse analysis*. London: Routledge. https://doi.org/10.4324/9780203200063

Crowell, A., & Kuhn, D. (2014). Developing dialogic argumentation skills: A 3-year intervention study. *Journal of Cognition and Development, 15*(2), 363–381. https://doi.org/10.1080/15248372.2012.725187.

Felton, M., Crowell, A., Garcia-Mila, M., & Villarroel, C. (2019). Capturing deliberative argument: An analytic coding scheme for studying argumentative dialogue and its benefits for learning. *Learning, Culture and Social Interaction*. https://doi.org/10.1016/j.lcsi.2019.100350

Fishman, E. J., Borko, H., Osborne, J., Gomez, F., Rafanelli, S., Reigh, E., Tseng, A., Million, S., & Berson, E. (2017). A practice-based professional development program to support scientific argumentation from evidence in the elementary classroom. *Journal of Science Teacher Education, 28*(3), 222–249. https://doi.org/10.1080/1046560X.2017.1302727

Ford, M., & Forman, E. A. (2006). Chapter 1: Redefining disciplinary learning in classroom contexts. In *, 30(1). Review of research in education* (pp. 1–32). https://doi.org/10.3102/0091732X030001001

Galton, M., Hargreaves, L., Comber, C., Wall, D., & Pell, A. (1999). *Inside the primary classroom: 20 years on*. London: Routledge. https://doi.org/10.4324/9780203269329

Geoffrey, N. L. (1983). *Principles of pragmatics* (1st ed.). London: Routledge. https://doi.org/10.4324/9781315835976

Gillies, R. M. (2015). Teacher dialogue that supports collaborative learning in the classroom. In L. B. Resnick, C. S. C. Asterhan, & S. N. Clarke (Eds.), *Socializing intelligence through academic talk and dialogue* (pp. 335–345). Washington, DC: AREA.

Hancher, M. (1981). What kind of speech act is interpretation? *Poetics, 10*(2–3), 263–281.

Hardman, J. (2019a). Developing and supporting implementation of a dialogic pedagogy in primary schools in England. *Teaching and Teacher Education, 86*. https://doi.org/10.1016/j.tate.2019.102908

Hardman, J. (2019b). Student talk in whole-class teaching: Findings from a teacher development intervention. In I. Meddegama, C. Danjo, J. Prudhoe, L. Walz, & R. Wicaksono (Eds.), *BAAL 2018: Taking risks in applied linguistics* (pp. 32–35). BAAL/Continuum.

Hardman, J. (2019c). Analyzing student talk moves in whole-class teaching. In N. Mercer, R. Wegerif, & L. Major (Eds.), *The Routledge International handbook of research on dialogic education* (pp. 152–166). New York: Routledge, Taylor & Francis Group.

Hennessy, S. (2020). Analytical coding schemes for classroom dialogue. In R. Kershner, S. Hennessy, R. Wegerif, & A. Ahmed (Eds.), *Research methods for educational dialogue* (pp. 95–130). London: Bloomsbury Academic. https://doi.org/10.5040/9781350060111.

Hennessy, S., Howe, C., Mercer, N., & Vrikki, M. (2020). Coding classroom dialogue: Methodological considerations for researchers. *Learning, Culture and Social Interaction, 25*, 1–19. https://doi.org/10.1016/j.lcsi.2020.100404

Hennessy, S., Rojas-Drummond, S., Higham, R., Márquez, A. M., Maine, F., Ríos, R. M., García-Carrión, R., Torreblanca, O., & Barrera, M. J. (2016). Developing a coding scheme for analyzing classroom dialogue across educational contexts. *Learning, Culture and Social Interaction, 9*, 16–44. https://doi.org/10.1016/j.lcsi.2015. 12.001

Heyd-Metzuyanim, E., Nachlieli, T., Weingarden, M., & Baor, R. (2020). Adapting a professional development program for cognitively demanding instruction across shifting contexts. *Educational Studies in Mathematics, 104*, 385–403. https://doi.org/10.1007/s10649-020-09967-y

Howe, C. (2009). Collaborative group work in middle childhood. *Human Development, 52*(4), 215–239. https://doi.org/10.1159/000215072

Howe, C., & Abedin, M. (2013). Classroom dialogue: A systematic review across four decades of research. *Cambridge Journal of Education, 43*(3), 325–356. https://doi.org/10.1080/0305764X.2013.786024.

Howe, C., Hennessy, S., Mercer, N., Vrikki, M., & Wheatley, L. (2019). Teacher–student dialogue during classroom teaching: Does it really impact on student outcomes? *Journal of the Learning Sciences, 28*(4–5), 462–512. https://doi.org/10.1080/10508406.2019.1573730

Howe, C., & Mercer, N. (2016). Commentary on the papers. *Language and Education, 31*(1), 83–92.

Howe, C., Tolmie, A., Thurston, A., Topping, K., Christie, D., Livingston, K., & Donaldson, C. (2007). Group work in elementary science: Towards organizational principles for supporting pupil learning. *Learning and Instruction, 17*(5), 549–563. https://doi.org/10.1016/j.learninstruc.2007.09.004

Jasniskaja, K., & Karagjosova, E. (2011). *Elaboration and explanation. Constraints in discourse IV*.

Keefer, M. W., Zeitz, C. M., & Resnick, L. B. (2000). Judging the quality of peer-led student dialogues. *Cognition and Instruction, 18*(1), 53–81. https://doi.org/10.1207/S1532690XCI1801_03

Kim, M. Y., & Wilkinson, I. A. G. (2019). What is dialogic teaching? Constructing, deconstructing, and reconstructing a pedagogy of classroom talk. *Learning, Culture and Social Interaction, 21*, 70–86. https://doi.org/10.1016/j.lcsi. 2019.02.003

Kershner, R., Hennessy, S., Dowdall, K., Owen, H., & Calcagni, M. E. (2020). Teachers as 'natural experimenters': Using T-SEDA to develop classroom dialogue. In L. Rolls, & E. Hargreaves (Eds.), *Reimagining professional development in schools* (pp. 102–122). London: Routledge. https://doi.org/10.4324/9780429293337.

Kuhn, D., & Udell, W. (2003). The development of argument skills. *Child Development, 74*(5), 1245–1260. http://www.jstor.org/stable/3696176.

Kunh, D., Wang, Y., & Li, H. (2011). Why argue? Developing understanding of the purposes and values of argumentive discourse. *Discourse Processes, 48*(1), 26–49. https://doi.org/10.1080/01638531003653344.

Kuhn, D., & Zillmer, N. (2015). Developing norms of discourse. In L. B. Resnick, C. S. C. Asterhan, & S. N. Clarke (Eds.), *Socializing intelligence through academic talk and dialogue* (pp. 77–86). Washington, DC: AREA. https://doi.org/10.3102/978-0-935302-43-1_6.

Larrain, A., Freire, P., López, P., & Grau, V. (2019). Counter-arguing during curriculum-supported peer interaction facilitates middle-school students' science content knowledge. *Cognition and Instruction, 37*(4), 453–482. https://doi.org/10.1080/07370008.2019.1627360

Littleton, K., & Mercer, N. (2013). *Interthinking: Putting talk to work* (1st ed.). New York: Routledge, Taylor & Francis Group. https://doi.org/10.4324/9780203809433

Lefstein, A., & Barak, M. (2020). Tradeoffs and dilemmas in coding classroom dialogue. In R. Kershner, S. Hennessy, R. Wegerig, & A. Ahmed (Eds.), *Research methods for educational dialogue* (pp. 130–134). Bloomsbury Academic.

Mehan, H., & Cazden, C. B. (2015). The study of classroom discourse: Early history and current developments. In L. B. Resnick, C. S. C. Asterhan, & S. N. Clarke (Eds.), *Socializing intelligence through academic talk and dialogue* (pp. 13–36). Washington, DC: AREA. https://doi.org/10.3102/978-0-935302-43-1_6.

Mercer, N., & Littleton, K. (2007). *Dialogue and the development of children's thinking*. New York: Routledge, Taylor & Francis Group. https://doi.org/10.4324/9780203946657

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. (2010). The analysis of classroom talk: Methods and methodologies. *British Journal of Educational Psychology, 80*(1), 1–14. https://doi.org/10.1348/00070990 9X479853

Michaels, S., O'Connor, C., & Resnick, L. B. (2008). Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in Philosophy and Education, 27*(4), 283–297. https://doi.org/10.1007/s11217-007-9071-1

Molinari, L., Mameli, C., & Gnisci, A. (2013). A sequential analysis of classroom discourse in Italian primary schools: The many faces of the IRF pattern. *British Journal of Educational Psychology, 83*(3), 414–430. https://doi.org/10.1111/j.2044-8279.2012.02071.x

Murphy, P. K., Wilkinson, I. A. G., Soter, A. O., Hennessey, M. N., & Alexander, J. F. (2009). Examining the effects of classroom discussion on students'comprehension of text: A meta-analysis. *Journal of Educational Psychology, 101*(3), 740–764. https://doi.org/10.1037/a0015576.

Nussbaum, E. M. (2011). Argumentation, dialogue theory, and probability modeling: Alternative frameworks for argumentation research in education. *Educational Psychologist, 46*(2), 84–106. https://doi.org/10.1080/00461520.2011.558816

Nystrand, M., Wu, L. L., Gamoran, A., Zeiser, S., & Long, D. A. (2003). Questions in time: Investigating the structure and dynamics of unfolding classroom discourse. *Discourse Processes, 35*(2), 135–198. https://doi.org/10.1207/S15326950 DP3502_3

O'Connor, M. C., & Michaels, S. (1993). Aligning academic task and participation status through revoicing: Analysis of a classroom discourse strategy. *Anthropology & Education Quarterly, 24*(4), 318–335. https://doi.org/10.1525/aeq.1993.24.4. 04x0063k

O'Connor, C., & LaRusson, M. (2014). Working with a low-inference discourse observation tool: What can we see?. In *Paper session presented at the meeting of Annual Educational Research Association, AREA, Philadelphia, PA*.

Osborne, J. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *Science, 328*(5977), 463–466. https://doi.org/10.1126/science.1183944

Osborne, J. F., Borko, H., Fishman, E., Gomez Zaccarelli, F., Berson, E., Busch, K. C., Reigh, E., & Tseng, A. (2019). Impacts of a practice-based professional development program on elementary teachers' facilitation of and student engagement with scientific argumentation. *American Educational Research Journal, 56*(4), 1067–1112. https://doi.org/10.3102/0002831218812059

Osborne, J. F., & Patterson, A. (2011). Scientific argument and explanation: A necessary distinction? *Science Education, 95*(4), 627–638. https://doi.org/10.1002/sce.20438

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

Pauli, C., & Reusser, K. (2015). Discursive cultures of learning in (everyday) mathematics teaching: A video-based study on mathematics teaching in German and Swiss classrooms. In L. B. Resnick, C. S. C. Asterhan, & S. N. Clarke (Eds.), *Socializing intelligence through academic talk and dialogue* (pp. 181–193). Washington, DC: AREA.

Peled-Elhanan, N., & Blum-Kulka, S. (2006). Dialogue in the Israeli classroom: Types of teacher–student talk. *Language and Education, 20*(2), 110–127. https://doi.org/10.1080/09500780608668716.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assesment Scoring SystemTM: Manual K-3*. Paul H. Brookes Publishing Co.

Pollak, I., Segal, A., & Lefstein, A. (2015). *Pedagogy in Israeli: Activity and discourse in classrooms*. In I. Pollak, A. Segal, & A. Lefstein (Eds.). Be'er Sheva: Laboratory for the Study of Pedagogy, Department of Education, Ben Gurion University of the Negev (in Hebrew).

Rapanta, C. (2019). *Argumentation strategies in the classroom* (1st ed.). Delaware: Vernon Press.

Resnick, L. B., Asterhan, C. S. C., & Clarke, S. (2015). *Socializing intelligence through academic talk and dialogue*. Washington, DC: AREA. https://doi.org/10.3102/978-0-935302-43-1

Resnick, L. B., Asterhan, C. S. C., & Clarke, S. N. (2018b). Accountable Talk: Instructional Dialogue that Builds the Mind. In *Educational Practices Series. The International Academy of Education (IAE) and the International Bureau of Education (IBE) of the Geneva, Switzerland: United Nations Educational, Scientific and Cultural Organization (UNESCO)*.

Resnick, L. B., Asterhan, C. S. C., Clarke, S. N., & Schantz, F. (2018a). Next generation research in dialogic learning. In G. E. Hall, L. F. Quinn, & D. M. Gollnick (Eds.), *The Wiley handbook of teaching and learning* (pp. 323–338). Hoboken: John Wiley & Sons, Inc.. https://doi.org/10.1002/9781118955901.ch13

Reznitskaya, A. (2012). Dialogic teaching: Rethinking language use during literature discussions. *The Reading Teacher, 65*(7), 446–456. https://doi.org/10.1002/TRTR.01066

Reznitskaya, A., Anderson, R. C., & Kuo, L. J. (2007). Teaching and learning argumentation. *The Elementary School Journal, 107*(5), 449–472. https://doi.org/10.1086/518623.

Reznitskaya, A., Kuo, L., Clark, A., Miller, B., Jadallah, M., Anderson, R. C., & Nguyen-Jahiel, K. (2009a). Collaborative reasoning: A dialogic approach to group discussions. *Cambridge Journal of Education, 39*(1), 29–48. https://doi.org/10.1080/030 57640802701952

Reznitskaya, A., Kuo, L., Glina, M., & Anderson, R. C. (2009b). Measuring argumentative reasoning: What's behind the numbers? *Learning and Individual Differences, 19*(2), 219–224. https://doi.org/10.1016/j.lindif.2008.11.001

Reznitskaya, A., & Wilkinson, I. A. G. (2017). *The most reasonable answer: Helping students build better arguments together*. Cambridge: Harvard Education Press.

Reznitskaya, A., & Wilkinson, I. A. G. (2021). The argumentation rating tool: Assessing and supporting teacher facilitation and student argumentation during text-based discussions. *Teaching and Teacher Education, 106*, Article 103464, 1–12 https://doi.org/10.1016/j.tate.2021.103464.

Rojas-Drummond, S., & Zapata, M. P. (2004). Exploratory talk, argumentation and reasoning in Mexican primary school children. *Language and Education, 18*(6), 539–557. https://doi.org/10.1080/09500780408666900

Searle, J. R. (1969). *Speech acts*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139173438

Segal, A., & Lefstein, A. (2016). Exuberant, voiceless participation: An unintended consequence of dialogic sensibilities? *L1 EducationalStudies in Language and Literature, 16*, 1–19. https://doi.org/10.17239/L1ESLL-2016.16.02.06

Shachar, H., & Sharan, S. (1994). Talking, relating, and achieving: Effects of cooperative learning and whole-class instruction. *Cognition and Instruction, 12*(4), 313–353. https://www.jstor.org/stable/3233683.

Shaffer, D. W., & Ruis, A. R. (2020). How we code. In A. R. Ruis, & S. B. Lee (Eds.), *Advances in quantitative ethnography* (pp. 62–77). Springer.

Song, Y., Chen, X., Hao, T., Liu, Z., & Lan, Z. (2019). Exploring two decades of research on classroom dialogue by using bibliometric analysis. *Computers & Education, 137*, 12–31. https://doi.org/10.1016/j.compedu.2019.04.002

TALIS 2008 Technical Report. (2010). Paris: OECD Publishing. https://doi.org/10.1787/9789264079861-en

Thurlings, M., & den Brok, P. (2017). Learning outcomes of teacher professional development activities: A meta-study. *Educational Review, 69*(5), 554–576. https://doi.org/10.1080/00131911.2017.1281226

Tschan, F., Zimmermann, J., & Semmer, N. K. (2018). Rules for coding scheme development. In E. Brauner, M. Boos, & M. Kolbe (Eds.), *The Cambridge handbook of group interaction analysis* (pp. 191–207). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316286302.011.

Vrikki, M., Wheatley, L., Howe, C., Hennessy, S., & Mercer, N. (2019). Dialogic practices in primary school classrooms. *Language and Education, 33*(1), 85–100. https://doi.org/10.1080/09500782.2018.1509988.

Walton, D. (2007). The speech act of clarification in a dialogue model. *Studies in Communication Science, 7*(2), 165–197.

Webb, N. M. (1989). Peer interaction and learning in small groups. *International Journal of Educational Research, 13*(1), 21–39. https://doi.org/10.1016/0883-0355(89) 90014-1

Webb, N. M., Franke, M. L., De, T., Chan, A. G., Freund, D., Shein, P., & Melkonian, D. K. (2009). 'Explain to your partner': Teachers'instructional practices and students'dialogue in small groups. *Cambridge Journal of Education, 39*(1), 49–70. https://doi.org/10.1080/03057640802701986

Webb, N. M., Franke, M. L., Ing, M., Wong, J., Fernandez, C. H., Shin, N., & Turrou, A. C. (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., Littleton, K., Dawes, L., Mercer, N., & Rowe, D. (2004). Widening access to educational opportunities through teaching children how to reason together. *Westminster Studies in Education, 27*(2), 143–156. https://doi.org/10.1080/0140672040270205

Weinberger, A., Marttunen, M., Laurinen, L., & Stegmann, K. (2013). *International Journal of Computer-Supported Collaborative Learning, 8*(3), 333–349. https://doi.org/10.1007/s11412-013-9173-4

Wells, G., & Arauz, R. M. (2006). Dialogue in the classroom. *Journal of the Learning Sciences, 15*(3), 379–428. https://doi.org/10.1207/s15327809jls1503_3

Wierzbicka, A. (1987). *English speech act verbs. A semantic dictionary*. Australia: Academic Press, Harcourt Brace Jovanovich.

Wilkinson, I. A. G., Murphy, P. K., & Binici, S. (2015). Dialogue-intensive pedagogies for promoting reading comprehension: What we know, what we need to know. In L. B. Resnick, C. S. C. Asterhan, & S. N. Clarke (Eds.), *Socializing intelligence through academic talk and dialogue* (pp. 37–50). Washington, DC: AREA. https://doi.org/10.3102/978-0-935302-43-1_3.