To appear in *The Journal of the Learning Sciences*

Teacher-Student Dialogue during Classroom Teaching: Does it Really Impact upon Student Outcomes?

Christine Howe, Sara Hennessy, Neil Mercer, Maria Vrikki, and Lisa Wheatley

Faculty of Education

University of Cambridge

Author Note

The research was conducted while all authors were affiliated to the Faculty of Education, University of Cambridge. Dr Vrikki has now moved to the Department of Education, University of Cyprus.

The research was supported (Grant: ES/M007103/1) by the Economic and Social Research Council of Great Britain (ESRC). The authors wish to thank the ESRC, the participating students, teachers and head teachers, and the large number of colleagues and postgraduates who assisted with project design, sample recruitment, data collection, and data preparation (especially Ayesha Ahmed, Annabel Amodia-Bidakowska, Sarah Baugh, Elisa Calcagni, and Helen Lancaster who helped with everything).

Correspondence about the research should be addressed to Professor Christine Howe, Faculty of Education, University of Cambridge, 184 Hills Road, Cambridge CB2 8PQ, UK. Email: cjh82@cam.ac.uk

Abstract

It is now widely believed that classroom dialogue matters as regards student outcome, with optimal patterns often regarded as requiring some or all of: open questions; elaboration of previous contributions; reasoned discussion of competing viewpoints; linkage and coordination across contributions; meta-cognitive engagement with dialogue; high student participation. To date however, the relevance of such features has been most convincingly examined in relation to small-group interaction amongst students; little is known about applicability to teacher-student dialogue. The paper reports a large-scale study that permits some rebalancing. The study revolves around the two lessons (covering two of mathematics, literacy and science) that were video-recorded in each of 72 demographically diverse classrooms (students aged 10-11 years). Key measures of teacher-student dialogue were related to six indices of student outcome, which jointly covered curriculum mastery, reasoning, and educationally relevant attitudes. Prior attainment and attitudes were considered in analyses as were other factors, e.g. student demographics and further aspects of classroom practice, that might confound interpretation of dialogue-outcome relations. So long as students participated extensively, elaboration and querying of previous contributions were found to be positively associated with curriculum mastery, and elaboration was also positively associated with attitudes.

Productive classroom dialogue

Around the end of the twentieth century, changes began to take place in the perspective that educational researchers adopted towards the verbal interactions that occur during classroom teaching. Prior to this, such interactions were typically regarded as following fixed patterns, prevalent amongst which was the so-called I-R-E/F format (Mehan, 1979; Sinclair & Coulthard, 1975), i.e. Initiation (e.g. 'What is the capital of Scotland?'), Response (e.g. 'Glasgow'), Evaluation/Feedback (e.g. 'No, it's Edinburgh'). Noting that Is and E/Fs were normally the prerogative of teachers, researchers focused on student Rs, with variation as a function of student gender, ethnicity and prior attainment widely studied (see Howe, 2010, for a review of this research). The changes in perspective amounted to the rejection of such patterns as fixed, and instead to their portrayal as options which, although frequently followed in practice, could in principle be superseded via alternative approaches. Furthermore, choice of option was not typically seen as neutral with respect to student outcomes: often with reference to 'socio-cultural theory' (e.g. Daniels, 2001; Wertsch, 1990), some approaches were spotlighted as especially productive as regards such outcomes. Without exception, these alternative approaches departed significantly from the I-R-E/F format, particularly where the latter involves the brief exchanges exemplified above.

Numerous proposals have been made about the characteristics that verbal interaction in classrooms should display in order to optimize student outcomes, but amongst these proposals five themes recur. The first is that rather than being restricted to closed questions like 'What is the capital of Scotland?', teacher initiations should also include open questions like 'Why has Edinburgh become a major centre of culture?', which permit multiple answers (e.g. Alexander, 2008; Mercer & Littleton, 2007; Nystrand, Wu, Gamoran, Zeiser, & Long, 2003; O'Connor, Michaels, & Chapin, 2015; Wells & Arauz, 2006). A second theme is that participants should make extended contributions, elaborating and building on previous

contributions made by themselves and others (e.g. Alexander, 2008; Boyd & Markarian, 2011; O'Connor et al., 2015; Rojas-Drummond, Littleton, Hernandez, & Zuniga, 2010; Wells & Arauz, 2006). A third theme is that differences of opinion should be acknowledged, probed and critiqued, ideally bringing in the reasons on which opinions are based (e.g. Alexander, 2008; Chinn & Anderson, 1998; Howe & Mercer, 2007; Lefstein, 2010; Mercer & Littleton, 2007; Mortimer & Scott, 2003; Osborne, Erduran, Simon, & Monk, 2001; Reznitskaya & Gregory, 2013; Schwarz, 2009; Wilkinson et al., 2017). A fourth theme is that through explicit links amongst contributions and attempts to co-ordinate (including when resolving differences), integrated lines of inquiry should be pursued (e.g. Alexander, 2008; Kumpulainen & Lipponen, 2010; Mercer & Littleton, 2007; Michaels, O'Connor, & Resnick, 2008; Osborne et al., 2001; Reznitskaya & Gregory, 2013). The final theme relates to the adoption of a meta-cognitive perspective upon verbal interaction, where participants should become aware of its value and reflect accordingly on their practice (e.g. Lefstein, 2010; Mercer & Dawes, 2008; Reznitskaya & Gregory, 2013; van der Veen, de Mey, van Kruistum, & van Oers, 2017). With the possible exception of the first theme, the expectation is that all participants should contribute fully to ensuring classroom interaction complies with these themes, implying high levels of involvement from students as well as teachers.

The themes emerged gradually after decades of classroom observation and teacher report. Unsurprisingly, therefore, there is considerable variation over how they are used. Researchers differ over which themes they emphasize, not necessarily because they reject the others, but probably more because of contrasting interests. Furthermore, they also vary over the labels used to integrate themes, especially over the term 'dialogue' (see Littleton & Howe, 2010). Some researchers use the term in a broad sense that encompasses all verbal interactions that occur during classroom teaching, i.e. both brief I-R-E/F and the supposedly more productive patterns. Such researchers typically qualify via terms like 'inquiry dialogue'

and ‘productive dialogue’ when addressing the latter (e.g. O’Connor et al., 2015; Wilkinson et al., 2017). Other researchers reserve ‘dialogue’ for the putatively beneficial patterns, often to link transparently with their concepts of ‘dialogic pedagogy’, which encompass a supportive ethos as well as verbalization. Dialogic ethos is typically seen to include such characteristics as open-mindedness, mutual respect, freedom from censure, reduced role division, and space to explore (e.g. Boyd & Markarian, 2011; Freire & Macedo, 1995; Haneda & Wells, 2008; Matusov, 2009; Wegerif, 2013). Finally, researchers differ over whether their approaches concur fully with the themes or imply refinement. For instance, while the forms of teacher questioning discussed in Chin (2007) mainly dovetail with the themes, they also include ‘verbal jigsaws’, which relate to the use of subject-specific technical terms. While Webb et al. (2009) spotlight reasoned utterances and hence the third theme, they also differentiate high- from low-level reasons depending on mathematical appropriateness.

Nevertheless, despite the heterogeneity, the five themes are indisputably prominent within the literature, endorsed by many researchers from many countries. Thus, for that reason alone, their relevance needs to be considered. Does their incorporation into classroom interaction genuinely boost student outcomes? Indeed, the question becomes even more pressing once its practical implications are recognized: through professional development programmes, many teachers are already being encouraged to implement the themes (see, e.g., Alexander, Hardman, & Hardman, 2017; Chinn, Anderson, & Waggoner, 2001; Haneda, Teenant, & Shearman, 2017; Hennessy, Dragovic, & Warwick, 2017; Lefstein & Snell, 2014; Pehmer, Gröschner, & Seidel, 2015; Pimentel & McNeill, 2013; Ruthven et al., 2017; Sedova, Sedlacek, & Svaricek, 2016; Wells & Arauz, 2006). It is therefore crucial to establish whether the encouragement is appropriate. This then is the rationale for the paper that follows, for the central concern is the relevance of the five themes (and the cross-cutting

theme of high student contribution) for student outcomes. The paper refers to the themes collectively as 'theoretically productive (classroom) dialogue', with 'productive' meaning 'conducive to positive student outcomes' and 'dialogue' employed in the broader of the senses discussed above but qualified via 'productive'. Using 'dialogue' in this fashion permits ready comparison with supposedly non-productive forms, and in any event with the focus on verbal interaction alone, linkage with 'dialogic ethos' is largely unnecessary. The key issue is whether as regards student outcomes theoretically productive dialogue is productive in practice as well as in theory.

Previous Research

There is already extensive research of relevance relating to the dialogue that occurs when students work independently of teachers in small groups. This research provides compelling evidence for the third of the above themes: the expression and reasoned discussion of differences of opinion has been shown repeatedly to support both academic attainment and general reasoning (e.g., Anderson, Howe, Soden, Halliday, & Low, 2001; Fung & Howe, 2014; Howe et al., 2007; Jurkowski & Hänze, 2015; Mercer, Dawes, Wegerif, & Sams, 2004; Mercer & Sams, 2006; Reznitskaya et al., 2009). For instance, in the studies of Anderson et al. and Fung and Howe, students worked together in small groups over ten lessons on the design of projects, subsequently writing individual reports about project implementation. Report quality was strongly predicted by the frequency of reasoned differences of opinion during group discussion. While obtaining parallel results with their extended *Thinking Together* programme, Mercer and colleagues also support the fifth theme: integral to their positive findings was the awareness of good practice achieved through negotiation and display of such 'ground rules' as 'We give reasons to explain our ideas' and 'If we disagree, we ask why'.

Perhaps unsurprisingly, the first theme (open vs. closed questions) has not featured prominently in small-group research, and there is also little evidence at present relating to the second theme, i.e. building on and elaborating (but see Rojas-Drummond et al., 2010). As regards the fourth theme, studies that (unlike the above) take groups out of class and assess their consequences in 'controlled' contexts frequently detect positive associations between referring back to preceding activities or dialogue and eventual student outcomes (for several such studies, see Howe, Tolmie, Duchak-Tanner, & Rattray, 2000). However, the relevance of reference back in natural settings is questionable (Howe et al., 2007). Yet whatever the case here, the fact remains that two themes have been strongly endorsed in the context of small-group activity amongst students, meaning that in this particular context patterns of dialogue do seem to matter. The trouble is that the context is far from mainstream as regards classroom teaching, being at best a minority enterprise but virtually unknown in some cultures (Howe, 2010). Even in the United Kingdom where group work has often received policy endorsement, research has documented students seated in small groups but seldom working collaboratively (and dialogically) around schoolwork (Kutnick & Blatchford, 2014). The prevailing dialogue in classrooms is between teachers and students, albeit between teachers and individual students, small groups or the whole class, and at present the impact of teacher-student dialogue on student outcomes requires further study.

This is not because of shortage of research addressing teacher-student dialogue. There have been hundreds of studies relating to this topic, and reviewing them, Howe and Abedin (2013) conclude that many are concerned with effectiveness. Some address dialogue in a fashion that does not map onto the present criteria for theoretical productiveness (e.g. Firestone & Brody, 1975; Hughes, 1973; Luckner & Pianta, 2011), but the majority are consistent. Whatever the case though, the modal approach according to Howe and Abedin is to analyze sampled dialogue qualitatively for compliance with models of effective practice,

whose appropriateness is presumed rather than tested. For that reason, student outcomes have seldom been assessed, let alone used in evaluation. Moreover, even when outcome data have been collected, their use is rarely compelling. In some cases, this is because dialogue was not analyzed and sometimes not even recorded, despite procedures that were implemented through dialogue (see, e.g., Adey & Shayer, 2015; Trickey & Topping, 2004). Elsewhere the problem lies with potentially confounding influences. For instance, there are studies that analyze teacher-student dialogue and assess student outcomes, and whose approach to dialogue concurs with the present one (e.g. Alexander et al., 2017; Herrenkohl, Palincsar, DeWater, & Kawasaki, 1999; Osborne, Simon, Christodoulou, Howell-Richardson, & Richardson, 2013; Ruthven et al., 2017). Sometimes (not always), these studies detect positive outcomes when theoretically productive forms are frequent. However, concerned with broader 'packages' that include good task design and/or productive small-group dialogue amongst students, these studies do not allow the effects of teacher-student dialogue to be isolated.

At the same time, it would be misleading to suggest that the impact of theoretically productive teacher-student dialogue is completely uncharted. Nystrand and colleagues (e.g. Applebee, Langer, Nystrand, & Gamoran, 2003; Nystrand, 2006) highlight the potential relevance to literacy attainment of the first of the above themes, i.e. that relating to open questions, and also of the crosscutting theme of high student contribution. Teacher-student dialogue that promotes student contribution has also been reported as having positive implications for attitudes to schooling and peers (Richter & Tjosvold, 1980) and for academic attainment in mathematics, science and literacy (Clarke, Xu, & Wan, 2010; Muhonen, Pakarinen, Poikkeus, Lerkkanen, & Rasku-Puttonen, 2018). Relating to the second and third themes, benefits for mathematics, science, literacy and oral communicative competence have been reported when teachers build on or elaborate students' ideas and encourage them to do

the same; or do such things while also encouraging students to explain and justify their ideas (Muhonen et al., 2018; O'Connor et al., 2015; van der Veen et al., 2017). Pauli and Reusser (2015) report positive associations between attainment in mathematics and what they term 'co-constructive talk', which includes explanation and justification. Yet while such studies are encouraging, they remain exceptional. Moreover, sample sizes can be small, the range of outcomes in each individual study is typically limited, and with the studies often addressing several aspects of dialogue simultaneously it is frequently unclear whether all or only some themes are supported. There is a pressing need for research that is both more extensive and more penetrating, and it was this need that motivated the study reported here. With an over-arching aim of supplementing current understanding about theoretically productive classroom dialogue, the present study examined the impact of dialogue in which teachers are involved upon three types of student outcome: academic attainment in mathematics, science and literacy, general reasoning, and educationally relevant attitudes. The effects of each examined aspect of dialogue were isolated from other aspects, permitting a highly nuanced picture of what is productive for each type of outcome.

Design Considerations

Insofar as previous studies have examined the relation between theoretically productive teacher-student dialogue and student outcomes, they have often employed interventionist (sometimes randomized controlled) methodologies. Attempts have been made to promote target features within intervention groups, and student outcomes have been compared with those achieved within control groups where the features were not promoted. Promotion usually involves workshops for teachers, and as such resembles the professional development programmes mentioned above. The trouble is that, no matter whether the focus is evaluation of theory, embedding in practice or both, the initiatives have achieved only patchy success as regards target dialogue (e.g. Alexander et al., 2017; Lefstein & Snell, 2014;

Osborne et al., 2013; Pehmer et al., 2015; Pimentel & McNeill, 2013; Ruthven et al., 2017; Wells & Arauz, 2006). Some features have proved more accessible than others, and even when features have emerged as accessible in principle teachers vary over their adoption. Moreover, when many teachers will have encountered claims about beneficial interaction independently of targeted interventions, it would be naïve to expect non-use of theoretically productive forms amongst control groups. In one study, average usage across control classrooms was actually higher than within the intervention group (Larrain, Howe, & Friere, 2018). Recognizing such difficulties, caution is already being advocated around the use of randomized controlled approaches in classroom research (e.g. Ginsburg & Smith, 2016).

As regards the study that follows, the message (from previous control groups) was that aspects of theoretically productive teacher-student dialogue will already be embedded in the practices of some classrooms and (from previous intervention groups) that the extent of embedding will vary across classrooms. This suggested naturally occurring variation with some features at least, whose relation to student outcomes might be examined without intervention. Certainly, this approach would have the merit of authenticity in that it would address existing practices rather than practices as artificially and possibly temporarily imposed. It would have the drawback of potentially having to discount features that were not widely used, but then as noted some features have proved resistant to intervention and therefore compromise this approach too. On balance, a naturalistic approach seemed preferable as regards the study, and this was the approach adopted. Teacher-student dialogue during routine lessons was analyzed for its approximation to what has here been presented as theoretically productive, i.e. for its compliance with the five specific themes (open questions, elaboration of previous contributions, reasoned discussion of competing viewpoints, linkage and coordination across contributions, meta-cognitive engagement with dialogue) and the cross-cutting theme of high student contribution. Analyzed dialogue was then used to address

the over-arching research question: are degrees of approximation to theoretically productive dialogue related to end-of-year student outcomes (with relevant start-of-year baseline performance considered)? At the same time, it was recognized that the approach was particularly susceptible to the problem highlighted already: the confounding influence of factors other than those being targeted. Accordingly, steps were taken to identify relevant confounds, and to take their influence into account.

Method

The study involved the teachers and students from 72 primary school classes, with data collection proceeding in three stages. The first stage was an initial visit to introduce the project, collect demographic and baseline data, and begin charting factors that might eventually be regarded as confounds. The second stage revolved around the video recording of teacher-student dialogue during routine lessons, while also supplementing data relating to potential confounds. The final stage focused upon the assessment of student outcomes.

Participants

Recruited through web, email and/or personal contact to teachers or head teachers, the 72 classes were spread across 48 schools. The schools were all located in England but covered London and the Home Counties (40%), East Anglia (34%), and Yorkshire and the West Midlands (26%). Urban and rural locations were represented. Twenty-eight schools supplied one class, 17 supplied two, two supplied three, and one supplied four. Class size varied from 20 to 36 students (M=27.76), and all classes were mixed-sex with between 19.23% and 79.17% girls (M=49.30%). Sixty-five classes comprised exclusively Year 6

students (aged 10 to 11 years), Year 6 being the final year of primary schooling in England, and seven classes were Year 5/6 composites. The classes were socio-economically and ethnically diverse (0-100% of students eligible for free school meals, *M*=19.3%; 0-96% from minority ethnic backgrounds, *M*=32.6%).

Procedure and Measures

After the institution's Ethics Committee had granted approval, the teachers were emailed comprehensive information about the study together with forms: a) for themselves to sign and return indicating willingness to participate throughout; b) for parents to sign and return indicating willingness for their child to participate in all (listed) aspects that went beyond routine teaching and assessment. The teachers were also sent a 'Teacher Questionnaire' (see later), which they were asked to complete and return electronically. Thereafter, the three stages of data collection began, being completed for each class between September and June in a single school year (2015-2016 for 27 classes; 2016-2017 for the remainder). The first stage (initial visits) was scheduled between early September and mid-October; the second stage (classroom recordings) took place between late October and late March; the third stage (student assessment) was completed between mid-May and early June.

Initial visits.

The visits began with informal briefing meetings between the researchers and class teacher/s, during which the teacher/s were asked to complete and/or return their consent forms if they had not already done this, and also to identify students whose parents had declined involvement in at least one aspect of the procedure. Thereafter, the researchers and teachers moved to the relevant classroom/s, and data collection commenced.

As noted, a major function of the data collected during the initial visits was the identification and assessment of factors that might confound interpretation of relations

between theoretically productive dialogue and student outcomes. Any factors that were associated with both dialogue and outcome would need to be treated as confounds, and their influence controlled. In planning the identification and assessment of such factors, it was recognized that more is known about the associates of outcome than about the associates of theoretically productive dialogue, with research relating to the latter more-or-less restricted to participant demographics (see Howe, 2010, for a review). Accordingly, literature relating to outcome operated as the starting point. Referring primarily to meta-analyses such as Hattie (2009), around 150 associates of student outcome were identified. The majority could be excluded a priori as of low effect size (e.g. composite vs. single-age classes), irrelevant for the sample (e.g. religious schools), or accounted for via other factors (e.g. low birth weight, which relates to academic attainment, but which is covered via assessment of immediate prior attainment). However, a priori exclusion seemed inappropriate with the 32 factors listed in the left-hand column of the Appendix, so all of these factors were assessed.

Details of how the 32 potential confounds were assessed are provided in the Appendix, but as regards the students they involved two instruments, both presented during the initial visits: a) NFER tests suitable for end-of-Year 5 or start-of-Year 6 (see www.nfer.ac.uk/schools/nfer-tests); b) a 'Child Questionnaire', which addressed six potential confounds, in four cases via inclusion of the Pupil Attitude to Self and School (PASS) survey (GL Assessment, 2013). PASS is a standardized instrument focusing on attitudes to school, self as learner, and relationship with teacher – see Appendix for sample items. Each student was asked to complete the NFER reading test *or* the NFER mathematics test (not both, to minimize demand). The two tests were divided at random within each class, usually through alternating tests across adjacently seated students (with the additional advantage of minimizing copying). The students were given 45 minutes to complete their test, working

under examination conditions. After a break (e.g. playtime), they then completed the Child Questionnaire. This was not given as a timed test but took around 20 minutes to complete.

If the teachers had not already completed the Teacher Questionnaire, they did this while the students were taking the NFER tests. As detailed in the Appendix, this questionnaire addressed 19 potential confounds, and focused upon the frequency with which specified practices occurred. The time devoted to the NFER tests also gave the researchers opportunities to note classroom conditions relevant for the subsequent recordings, e.g. levels of illumination, background noise. Sketches were made to show the positions of tables, computers, windows and doors, and the numbers of students around each table.

Classroom recordings.

. Each recording session involved two researchers, one whose prime responsibility lay with video-recording teacher-student dialogue and one whose focus was observing small-group activity. When the study's central concern was the former, the established influence of the latter was recognized as a potential confound. Video recording involved a camera (attached to a tripod) placed in an unobtrusive area of the classroom, and two microphones to ensure high sound quality, one attached to the teacher and the other for the ambient sound. Students whose parents had declined involvement in recordings were taken out of class or seated out of camera range. With everything ready, the teachers were asked to proceed as normal, the class was encouraged to ignore the camera, and about 30 minutes familiarization footage was recorded and subsequently discarded. Data collection proper began with the onset of a lesson in mathematics, science or literacy, with onset easily recognized from teacher announcement. Recording continued until, again from teacher announcement, it was clear that the target lesson had ended, and another subject was to be studied or a break was to take place. Lesson duration varied from 30 to 102 minutes (M=65.40 minutes, SD=14.23).

Two or three lessons were recorded in each classroom,[1] although to ensure comparability across classes, only two lessons per class were included in the study, even when three were recorded. These two lessons always covered different subjects from mathematics, science and literacy, with selection of lessons otherwise randomized when three had been recorded. This resulted in a final sample where recordings covered mathematics and science in 15 classes, mathematics and literacy in 43 classes, and science and literacy in 14 classes. The two lessons were recorded on the same or successive days, for the distances that the researchers often had to travel precluded spaced visits. While this meant that the lessons were clustered at specific points between October and March, there was no reason to anticipate lack of comparability due to associations between dialogic practices and time of recording. By October, the students would have settled with their teachers and potentially disconcerting preparations for end-of-year assessment typically escalate after March.

Whenever the teacher asked the students to work independently in small groups during the recorded lessons, the relevant researcher chose one group at random and, following procedures adapted from Howe et al. (2007), assessed the quality of the group's interaction. Employing time sampling techniques (10-seconds per minute preparation, 20-seconds observation, 10-seconds recording on grids, 20-seconds rest), the researcher noted usage of key features of dialogue, such as elaboration, disagreement, and justification. Noted features were then used to rate the group on five 3-point scales (1=*Not true of the group*, 2=*Partly true*, 3=*Very true*). A further five scales were used to rate more general aspects of group functioning (see Appendix for sample scales). The two researchers who, across the classes, observed group work had been trained following Howe et al.'s rigorous procedures

[1] Analyses conducted when two-thirds of the classrooms had been visited indicated that dialogue indices computed from two lessons almost perfectly predicted indices computed from three, so long as the lessons covered different subjects (i.e. 90-95% of the variance was covered). Thus, to expedite data collection and coding, two lessons only were recorded thereafter.

and had achieved 81% agreement over their independent ratings during training. In addition to the ratings, the researchers also noted general features, e.g. number of students in the group, gender composition, whether the overall style approximated 'collaboration' or 'peer tutoring' (Damon & Phelps, 1989). The researchers stayed with the same group over very short sequences of group work punctuated by dialogue with teachers, and their ratings applied across the whole sequence. Otherwise, they moved to different groups for each group-work session in the interests of representativeness.

Student assessment.

Within the socio-cultural tradition that, as noted, underpins much research into classroom dialogue, apparent progress in students' learning or understanding during dialogue itself is often taken as indicative of positive student outcomes. However, such progress does not necessarily generalize to other contexts or relate to longer-term growth (e.g. Howe, 2009; Howe & Zachariou, 2017). To optimize the present study's power, it seemed preferable to target generalized, longer-term outcomes directly, and therefore assess the students individually some months after the classroom recordings. In the event, a six-component approach to assessment was followed, with three components relating to the Standard Assessment Tests (SATs) that most Year 6 students in England take during May each year. SATs are government-prepared, teacher-administered, timed tests, which address mastery of prescribed curricula in: a) Mathematics; b) Grammar, punctuation and spelling (commonly known as SPAG); c) Reading (see www.gov.uk/government/publications - STA/16/7907/e. The results have major implications for primary schools since performance tables are published, with consequences for school reputation and resourcing. In accordance with UK Government policy, the SATs taken during May 2016 (i.e. the study's first year) were more challenging than previously, but difficulty levels were constant across the study's two years.

Recognizing the socio-political significance of SATs,[2] the 72 teachers in the sample were asked to supply their students' individual scores (suitably anonymized) for all three SATs. Sixty-eight teachers provided data, and with Reading and SPAG all scores were usable (*N*=1751 and 1754 students respectively). However, for a proportion of the students in some classes, it was not the class teacher who taught mathematics, but a colleague. Since the anonymization meant that these students could not be differentiated from their classmates, it was decided to discount the mathematics data relating to their whole class. Thus, analyses relating to the Mathematics SAT are based on 61 classes (1573 students).

There had once been a SAT in science too, but by the time of the study this was no longer the case. Therefore, knowledge and understanding in science were assessed via a specially prepared test. The test was designed in collaboration with primary science specialist teachers, was restricted to material that the statutory curriculum specifies for Year 6, and had conceptual and procedural components. Both components used multiple-choice and short-text items. The conceptual component addressed inheritance and evolution, this being the curriculum-specified topic that more teachers anticipated covering than any other when asked about their plans via the Teacher Questionnaire. Test items covered fossilization, heritability (including of acquired characteristics), and inter-generational adaptation. Many items were adapted with only minor changes from sample assessments that the UK Government has published; the remainder were grounded in research literature relating to the age group. The procedural component revolved around the recognition and design of fair tests, and the drawing of conclusions from test data. Items were sourced once more from the government's sample assessments and the research literature. Pilot data indicated that the items had

[2] The desirability of including outcome measures that policymakers and hence schools regard as highly significant was a major reason for the study's focus upon Year 6.

excellent scale properties (Cronbach's alpha across all items=.95; mean square out-fit scores all below 2.00 on Rasch analysis).

In addition to the specific curriculum subjects, it seemed desirable to evaluate general reasoning. As noted, small-group dialogue amongst students has been found to impact positively on reasoning, so the same might apply with teacher-student dialogue. Yet while there are many published tests of reasoning, few are suitable for the study's age group, so once more a specially designed test was employed (Ahmed, Howe, Major, Hennessy, Mercer, & Warwick, Forthcoming). The test covered: a) facts and opinions, e.g. indicate whether 'The weather in Britain is awful' is a fact or an opinion; b) reasons and conclusions, e.g. underline a sentence that contains a conclusion in 'Teaching is the best job in the world. You get to know lots of young children and help them learn new things…' c) saying and implying, e.g. 'The explorer says "To survive we must drink this water" What does the explorer imply?' - options = 'The water is clean', 'We can't survive without drinking the water', 'The water is delicious'; d) comparison of reasons, where two characters give reasons for some viewpoint, e.g. about keeping children in at lunch-time when they misbehave, and the task includes judging whether one reason repeats or differs from the other, or is relevant or irrelevant to the viewpoint. Once more, pilot data indicated excellent scale properties (Cronbach's alpha across all test items=.98; two mean square out-fit scores just above 2.00 on Rasch analysis, but otherwise all below).

The final assessment component was a second presentation of PASS, this time to examine attitudes close to the school year's conclusion. As regards administration of the science and reasoning tests and PASS, the teachers were offered the options of researcher- or teacher-presentation. All but five opted for teacher-presentation, so materials were normally despatched to the schools shortly after completion of SATs with administration procedures detailed. These included restricting the tests to Year 6 students in the composite classes, and

treating science and reasoning as timed tests (35 and 30 minutes respectively) while not placing time restrictions on PASS. If the teachers had previously reported covering inheritance and evolution, all three assessments were to be completed, ideally on the same day but in any event close in time. Order of presentation was to be Reasoning-PASS-Science or Science-PASS-Reasoning (instructions varied to achieve balance across classes). If the teachers had not covered inheritance and evolution, the science test was to be omitted, and the order was to be Reasoning-PASS. In the event, the reasoning test and PASS were completed in all 72 classrooms (respectively 1778 and 1784 students, while the science test was completed in 44 classrooms (1103 students).

Data preparation

The study's aim required examination of the relations between theoretically productive teacher-student dialogue and the six outcome measures while also considering relevant prior performance and avoiding distortion from the potential confounds. As realizing the aim depended upon parametric analyses, the challenge was to create meaningful dialogue, outcome, baseline, and confound variables consistent with such analyses.

Dialogue variables.

Video-recorded dialogue was analyzed using a scheme based on work reported by Hennessy, Rojas-Drummond and colleagues (Hennessy et al., 2016). This work coupled a comprehensive review of relevant studies with an attempt to develop categories that represent and synthesize the conceptualizations of productive dialogue informing the studies. Both aspects of Hennessy et al.'s work were referred to when identifying the themes that underpin the present research, i.e. open questions, elaboration of previous contributions, reasoned discussion of competing viewpoints, linkage and coordination across contributions, meta-cognitive engagement with dialogue, and cutting across these, high levels of student

contribution. Evaluation and sometimes modification or supplementation of Hennessy et al.'s representational categories in accordance with the identified themes resulted in the scheme used here to analyze the recorded dialogue. Thus, theme identification and scheme development proceeded in an intertwined fashion from shared roots, guaranteeing synergy between them (see also Vrikki, Wheatley, Howe, Hennessy, & Mercer, Forthcoming). Application of the scheme for purposes of analysis was restricted to sequences involving teachers (albeit with the whole class, small groups or individual students), but within those sequences contributions from teachers and students were both examined. With the earliest 86 lessons amongst the set of 144, the recorded dialogue was transcribed professionally, and analyzed using both transcripts and video. With the analytic scheme then familiar, the remaining lessons were analyzed directly from the videos.

Turn-level codes. The analytic scheme required that each turn (identified via speaker switch) be coded as detailed in Table 1. Of the codes, ELI, REI and CI represented open questions, so were relevant to the first of the key themes. Addressing elaborating and building upon, EL related to the second theme, as did ELI. Q and RE (and REI) were concerned with difference and reasoned justification, and therefore bore on the third theme. The fourth theme of linkage and integration was examined via SC, RC, RB and RW, with CI also relevant. When RC was detected, its constituent reason was not also coded RE, i.e. RC took priority over RE. OI and UC were included not because they were theorized as productive but because they allowed theoretically productive forms to be evaluated relative to other forms. Code A was likewise not theorized as productive, and since it could in principle co-occur with all codes bar Q (and did so weakly in practice, Pearson *r*s=.14 to .44) it was viewed as uninformative as regards this study. It is not considered further.

-Insert Table 1 about here-

If multiple codes applied with a single turn they were all noted, but multiple instances of a single code within one turn were only noted once. Thus, when one teacher said (replying to 'Only one person speaks at a time' from a student), 'Well done, we should always remember that in a good conversation, one person should be speaking at a time. So we have seven groups today, we should only be able to hear seven voices', her turn was coded A (for 'Well done') and EL for the remainder even though there were actually two instances of elaboration. To check reliability, two researchers independently coded 12 lessons before transcript/video-based coding began, and a further eight lessons before moving to video-only coding (9952 turns in total). Excluding CI and RC, which were seldom detected, correlations (Pearson *r*s) over the frequencies per lesson of each code ranged from .65 to 1.00 *(M*=.79), all *p*s<.002[3].

The route from coded turns to parametric variables began with the frequencies for each code across the two lessons recorded in each classroom being added together. Then, in order to correct for the varying lengths of lesson pairs, the frequency totals were divided by the total duration of the two lessons in minutes. To produce realistic values, corrected frequencies were also multiplied by 130.81, i.e. a constant representing the average duration in minutes of lesson pairs across all 72 classes. All analyses reported subsequently employ these values, with their means (and variability) across classes indicated in Table 1. As the table shows, the rarity of CI and RC detected during the reliability checks was confirmed with the full sample, and the frequency of SC was also extremely low. As a result, there seemed little point in taking the co-ordination codes further. By contrast, three types of invitation (ELI, REI and OI), two types of statement (EL and RE), and UC (often short

[3] As noted, the background research literature and associated professional development emphasize boosting the *frequencies* with which theoretically productive dialogue is used. Therefore, to address this background, the analyses to be reported later are necessarily also based on frequencies. For this reason, correlations between totals are the appropriate indicators of reliability rather than turn-based indicators like kappa or Krippendorf's alpha.

replies to OI) were ubiquitous. Moreover, principal components analysis[4] of the frequencies (Oblimin rotation with Kaiser normalization; 83% of the variance explained) indicated that these six codes fell into three clusters: a) Elaborated – ELI loaded .92, EL loaded .94, other loadings -.05 to .08; b) Reasoned – REI loaded .90, RE loaded .94, other loadings -.14 to .13; c) Non-Dialogic – OI loaded .89, UC loaded .82, other loadings -.06 to .06. This suggested combination along cluster lines, and as cluster frequencies for teacher and student contributions combined were strongly correlated with cluster frequencies for each taken separately (Pearson *r*s=.88 to .96 for teachers; .85 to .93 for students), combination across participants was also suggested.

Accordingly, two variables adopted for purposes of analysis were Elaborated and Reasoned. Respectively based on EL combined with ELI and RE combined with REI, contributions exemplifying these variables appear in the top four rows of Table 1. Two further variables were the Elaborated/Non-Dialogic Ratio, and the Reasoned/Non-Dialogic Ratio, computed such that *low* ratios signaled high frequencies of Elaborated and Reasoned relative to Non-Dialogic. Jointly, these four variables allowed examination of absolute frequencies and frequencies relative to theoretically non-productive forms. All four could in principle occur with moves typifying the remaining codes (Q, RB, RW, also exemplified in Table 1), and this may be why their frequencies were never more than weakly correlated with the frequencies of Q, RB and RW (Pearson *r*s=-.09 to .30). For that reason, Querying, Referring Back, and Referring Widely were preserved as separate variables, making seven turn-based variables in total.

[4] Data reduction across this paper is intended merely to simplify measured variables, so following Field (2013) principal components analysis was regarded as appropriate throughout. However, noting controversy within the literature over techniques, all relevant analyses have been repeated using factor analysis. Reported results were always replicated, e.g. here the same, strong, three-way solution emerged with no cross-factor loadings.

Frequency distributions across the 72 classes indicated that all seven variables were suitable for inclusion in parametric analyses, but they were of course also highly abstract, and this raised concerns about meaningfulness. Accordingly, five transcripts were chosen that varied over the absolute frequencies of Elaborated and Reasoned and also over the two Ratios. These transcripts were sent without disclosing selection criteria to eight leading scholars in the field, who were asked to rank them for the extent to which, from the scholar's own perspective, they represented educationally productive dialogue. All eight scholars had previously endorsed most (often all) of the themes, meaning that their rankings bore directly on theme-variable coherence. Some scholars completed the exercise themselves, while others involved their teams. The mean rank order across scholars correlated (Spearman rho) .76 with the rank order predicted by the Ratios but only .23 with the rank order predicted by the absolute frequencies. The implication is that by virtue of the Ratio variables, the turn-based variables did discriminate amongst recorded lessons in a fashion that the research community would find meaningful. Importantly, this is true regardless of whether the Ratio variables are actually predictive of student outcomes.

Lesson-level ratings. In addition to the turn-based coding, each lesson was rated holistically for five relatively strategic aspects that could not readily be captured at the turn level: a) Aims, covering lesson aims and objectives; b) Monitoring, addressing the monitoring and guidance of student activity; c) Reflection, relating to how the learning process was dealt with; d) Talk Rules, concerned with the coverage of productive dialogue (as with the above 'Well done etc.' example); e) Student Participation, assessing opportunities for students to express ideas and engage with the ideas of others. Three-point scales were used: 0=*Aspect non-evident during the lesson*; 1=*Aspect evident but teacher-led*; 2=*Aspect evident and student input*. Monitoring, Reflection and (especially) Talk Rules were regarded as addressing the fifth, meta-cognitive theme from those listed earlier, while all five

scales (but especially Student Participation) were seen to bear on the crosscutting theme of student contribution.

At the time the above five scales were applied, four further 3-point scales, adapted from Danielson (1996), were also used, this time to assess additional potential confounds (see Appendix). Reliability checks based on the independent ratings produced by two researchers indicated agreement that ranged from 67% to 100% across the nine scales, with mean agreement of 81%. Taking each scale separately, the ratings for the two lessons recorded in each classroom were added together to produce nine scores per class, each between 0 and 4. Combination by simple addition was justified when each lesson made an identical contribution to the total, i.e. a value of 0, 1 or 2. With eight scales, the distribution of scores across classes showed suitability for parametric analysis, so the totals out of 4 became the subsequent variables. Talk Rules however produced highly skewed data, due to 89% of lessons being rated 0. While this may well have reflected scant usage, it is also possible that rules had been negotiated earlier in the life of a class, and no longer required explicit discussion. Either way, concerns about validity resulted in data from the Talk Rules scale being discounted. As regards the other four dialogue-related scales, principal components analysis provided modest grounds for combining Monitoring and Student Participation, but no grounds for combining Aims or Reflection with any other scale. Accordingly, it seemed safest to keep all four scales separate, meaning that the eventual dialogue variables amounted to seven derived from turn-level codes and four derived from lesson-level ratings. Table 2 presents the distributional properties of these variables across the 72 classrooms.

- Insert Table 2 about here -

Outcome variables.

Distributional data relating to the outcome variables are also shown in Table 2. As regards the three SATs (Mathematics, SPAG, Reading), most teachers supplied scores in

standardized form, meaning that each student's scores lay between 80 and 120. When raw scores were supplied, these were standardized in-house. The maximum scores possible with the science and reasoning tests were respectively 33 and 40. With the science test, a scoring manual had been prepared for assessing responses to the short-text items, and markers were trained in its application. Their reliability was checked through comparing their individual scores for these items across the students in eight classes: inter-judge agreement was 92% for all pairs of markers. With PASS, each of the 50 items was scored from 1 to 4 (see Appendix), and scores for negatively worded items were reversed so that 4 always indicated a maximally positive attitude. Cronbach's alpha across all items was .93. Discounting the six students who omitted more than 5% of items, individual mean scores were computed across the 50 items (so maximum score possible=4). With no subjective element to the multiple-choice items in the science test or any items in the reasoning test and PASS, formal reliability checks were unnecessary. Nevertheless, 10% of scripts from all classes were subject to quality checks, with fewer marking errors detected than one per 500 answers.

To ensure that dialogue analysis was 'blind' to student outcome, separate teams completed these two aspects of data preparation. Indeed, until data preparation was complete, only the project administrator had access to collated scores. Equivalent precautions (over marker training, quality control and blinding) were taken with the scoring of those elements from the initial visits to be used as start-of-year baselines, specifically: a) Class means across NFER Mathematics scores (hereafter PreMaths (M)) for Mathematics SAT; b) Class means across NFER Reading scores (hereafter PreRead (M)) for SPAG and Reading SATs; c) Class means across NFER Mathematics and Reading scores combined for science and reasoning; d) Class means across start-of-year PASS scores (hereafter PrePASS) for PASS when used as an outcome measure. PreMaths (M), PreRead (M) and PrePASS were also potential confounds, i.e. they had the dual function of providing baseline data for some outcome variables and,

potentially at least, confounding interpretation of dialogue-outcome relations due to associations with both. As potential confounds, their scoring is detailed in the Appendix.

Potential confounds.

In fact, the Appendix outlines the scoring of data relating to all potential confounds, and (in the third column) lists the 33 variables that were derived. As indicated, all confound variables were computed as class-level measures, even when they were based on responses from individual students. This was because the students only provided comprehensive, individual data with four variables (PrePASS, PI_Talk, PI_Visit and Mobility). When the remaining 29 variables were either assessed at the class level in the first place or necessarily class averages due to derivation from student sub-samples, individual-level representation for four variables alone seemed an unnecessary complication. As the Appendix shows, the eventual variables did not lie precisely in one-to-one correspondence with the potential confounds. First, there were four indices of prior attainment, PreMaths (M) and PreRead (M) and, to reflect spread, their respective standard deviations. Second, the high Cronbach's alpha for PrePASS (.93 as with the end-of-year data) suggested that a composite score would suffice to represent the four associated factors: separate scores for each potential confound were unnecessary. Third, principal component analysis of responses to the Child Questionnaire items covering parental involvement (Oblimin rotation with Kaiser normalization; 47% of the variance explained) indicated two groups: a) items covering *discussion* about school work (relevant items loaded .44 to .79; other loadings -.09 to .14); b) items covering *visits* to school (relevant items loaded .46 to .71; other loadings -.01 to .08). Thus, two indices of parental involvement were computed.

Group work was not observed during eight of the 144 lessons. These lessons were awarded ratings of 1 on each of the ten scales, but otherwise the actual ratings were used with average ratings derived for each scale across the group work sessions observed in each

lesson. As Cronbach's alpha across these averages was very high (.90), it seemed reasonable to use the mean rating across the 2 (lessons) x 10 (scales) as the quality index for each classroom. Importantly, all observed group work was collaborative in nature, i.e. students worked jointly without prescribed roles; peer tutoring was never observed. Thus, peer tutoring could be excluded from the battery of confound variables without further consideration. Exclusion was also justified with five variables where virtually all teachers indicated the supposedly optimal situation: Homework, *M*=4.92, *SD*=0.33; Exercise, *M*=4.97, *SD*=0.16; IWB, *M*=4.82, *SD*=0.48; Computers, *M*=4.82, *SD*=0.39; PropFluent, *M*=97.09%, *SD*=7.71 (see Appendix – column headed 'Sample variation'). Twenty of the remaining 27 variables displayed distributional properties across the 72 classes that were consistent with parametric analysis. With five of the exceptions, the problems could be fixed via log transformation or, with only one or (once) two outliers, winsorizing (Field, 2013). There was no obvious solution with the other two variables, Thinking and %NoESL, so although not excluded these variables were, as explained later, treated as special cases.

Results

Data analysis revolved around the relationships amongst the 11 dialogue variables, the 27 remaining confound variables, and the six outcome variables. However, before examining these relationships, two further checks were made on the dialogue variables themselves. The expert ratings had shown them to be reasonable proxies for theoretically productive dialogue, but could they also be taken as stand-alone indices of practice? As noted, all classrooms were recorded at one point between October and March, and while associations between time of recording and patterns of interaction were considered unlikely, it seemed prudent to check. Accordingly, correlations (Pearson *r*s) were computed between the number of days after 1st October that each classroom was recorded and each of the

dialogue frequencies. All correlations were non-significant (*M*=.02, range=-.11 to .15). In addition, there were instances where several classes were drawn from the same school, and while school influences upon classroom dialogue seemed implausible, once more checking was desirable. Regression analyses indicated that school effects were negligible: with eight dialogue variables, the proportion of variance covered by school effects was no more than 0.2%, and the mean proportion across all 11 variables was 1.20%, range=0.1% to 5%.

With the above as backcloth, relationships were analyzed following two steps. The first step was to establish which, if any, of the potential confounds to include when examining dialogue-outcome relations, i.e. which potential confounds were actual confounds. Variables extrinsic to dialogue-outcome relations only distort interpretation of those relations when they are associated with both dialogue and outcome. Thus, the key issue was whether any confound variables could be discounted because they were not associated with either dialogue or outcome. Examining this issue involved two sets of analyses, one assessing confound-dialogue relations and the other assessing confound-outcome relations. With the genuine confounds identified, the second step was the analysis of dialogue-outcome relations per se, with the effects of confounds controlled. All analyses were conducted using the *Statistical Package for the Social Sciences* Versions 22-25 (SPSS, Chicago, IL, USA), as indeed were the analyses reported above. Interaction effects were interrogated via PROCESS (Preacher & Hayes, 2004), downloaded into SPSS.

Preliminary analyses

With dialogue and confound variables both measured at the class level, confound-dialogue relations could appropriately be examined using multiple regression. Accordingly, 11 backward elimination regression analyses were conducted with each dialogue variable as the dependent variable, and the potential confounds as the independent variables. The analyses were actually conducted twice, once assessing the 25 non-problematic variables and

once with the problem cases (Thinking and %NoESL) included. With 11 analyses and 25 or 27 independent variables, there was a high probability of significant associations arising by chance. Recognizing this, it seemed reasonable to eliminate confound variables when their associations with at least 10 dialogue variables were statistically non-significant and their association with any eleventh was only significant at the .05 level.[5]

Following these criteria, 18 potential confounds were eliminated (see Appendix, column headed 'Dialogue related'), including all variables relating to participant demographics. Of the confound variables that remained, those relating to prior attainment featured prominently: a) PreMaths (M) predicted the frequencies of Elaborated and Querying, the Elaborated/Non-Dialogic Ratio, and the ratings for Monitoring (ßs=.12 to -.46, *p*s=.046 to .002); b) PreRead (M) predicted the frequencies of Elaborated and Reasoned, the Reasoned/Non-Dialogic Ratio, and the ratings for Reflection (ßs=-.25 to .49, *p*s=.037 to <.001); c) PreMaths (SD) predicted the frequency of Querying and the ratings for Reflection (respective ßs=.29 and .25, *p*s=.005 and .032). Also prominent were confound variables derived from video ratings: d) Goal clarity (Goals) predicted the frequencies of Querying, Referring Back and Referring Widely and the ratings for Aims and Monitoring (ßs=-.13 to .58, *p*s=.032 to <.001); e) Quality of feedback (Feedback) predicted the frequency of Reasoned, the Reasoned/Non-Dialogic Ratio, and the ratings for Monitoring and Student Participation (ßs=.25 to .96, *p*s=.017 to <.001); f) Quality of teacher-student relations (Relations) predicted the frequency of Querying (ß=-.29, *p*=.009). These associations are perhaps unsurprising when scale usage would have partially depended on dialogue, some scales relate to the concept of 'dialogic pedagogy' discussed earlier, and some relate conceptually to the dialogue variables themselves, e.g. Goals and Aims. Finally: g) PrePASS

[5] Two-stage linear step-up procedures (Benjamini, Krieger, & Yekutieli, 2006) confirmed that this approach resulted in no inappropriate exclusions.

predicted the two Ratio measures and the frequency of Referring Widely (ßs=-.22 to -.28, *p*s=.047 to .018); h) Quality of collaborative group work (Group) predicted the frequency of Elaborated and the ratings for Student Participation (respective ßs=.23 and .31, *p*s=.045 and .002); i) the frequency of school trips (Trips) predicted the frequency of Elaborated (ß=.29, *p*=.009).

As regards confound-outcome relations, there is of course extensive research attesting to strong associations involving every potential confound. Yet this does not necessitate associations within the present sample, so here too the issue was examined empirically. This time two-level modelling was used, recognizing that the covariates (the nine remaining potential confounds) were measured at the class level, while the dependent variables (each of the six outcome variables) were measured at the student level with the students clustered in classes. Of course, some classes were also clustered in schools. However, the negligible school effects meant that two-level modeling (class, student) was sufficient, with the sample size also more than adequate for this approach (Field, 2013; Maas & Hox, 2005). PreMaths (M), PreRead (M), and PrePASS were excluded when the analyses addressed outcome variables for which they had been designated baseline indices of prior performance (see earlier). These variables would be included in the main analyses in any case. In the event, the three variables derived from video ratings (Goals, Feedback, Relations) were never significantly related to student outcome, and nor was Trips. However, as Table 3 indicates PreMaths (M), PreMaths (SD), PreRead (M), PrePASS and Group were all significantly related to at least one outcome variable (see also Appendix - column headed 'Outcome related'). Thus, as regards the main analyses, these five potential confounds needed to be treated as actual confounds, and included in the analyses wherever the consequences of dialogue variables they also related to were examined.

- Insert Table 3 about here -

Main analyses

Dialogue-outcome relations were also examined via two-level (class, student) modelling, regarding this as the optimal strategy for relating dialogue (assessed at the class level) to the scores of individual students while also recognizing that the students were clustered in classes and school effects were immaterial. The main effects of each dialogue variable were examined separately. Then interactions were analyzed between each of Querying, Referring Back and Referring Widely and each of Elaborated, Reasoned and the two Ratio measures. Given the themes outlined earlier, it was felt that the meaning of the first three variables might depend on their association with the other four. Finally, interactions were assessed between Student Participation and each of the other variables, recognizing the former as the major index of student contribution, i.e. the crosscutting theme. Thus, every model included one or two dialogue variables as covariates, with variables (and interaction terms) centered when examining interactions. Also included as covariates were the baseline indices, together with confound variables as indicated via the preliminary analyses (see Table 4). When interactions were examined, the covariates associated with both dialogue variables were included. For instance, examination of the interactive impact of the Elaborated/Non-Dialogic Ratio and Student Participation on SPAG scores required the inclusion of PreRead (M) (the baseline index for SPAG), PreMaths (M) and PrePASS (confounds due to their associations with the ratio measure and SPAG) and Group (confound due to its association with Participation and SPAG). While statistical significance was set at the conventional level of $p<.05$ (two-tailed), dialogue-outcome relations associated with p values between .05 and .1 are also reported below.[6] This is partly to avoid missing trends, but mainly because the

[6] In the interests of brevity, dialogue-outcome relations are not reported where $p>.1$, and the effects of confound and baseline variables are also omitted. Full models will be supplied upon request, as will analyses reported in summary form elsewhere in the paper.

inclusion of variables associated with both dialogue and outcome, i.e. the confounds, means that dialogue-outcome relations are consistently underestimated (Miller & Chapman, 2001).

- Insert Table 4 about here -

Two-level models revealed no statistically significant main effects of dialogue with any of the three SAT scores, and no significant interactions relating to Reading. However, as detailed in Table 5, three interactions were statistically significant with Mathematics, and two were statistically significant (or very close) with SPAG. Four of these interactions involved Student Participation, and as clarified in Figure 1 they all signal that when levels of Student Participation were high, high levels of Elaborated and Querying were productive. With low levels of Student Participation, high levels of Elaborated and Querying were not productive. Moreover, when levels of Elaborated and Querying were low, high levels of Student Participation were also largely irrelevant. The fifth interaction (presented in Figure 2 with others relating to Referring Back) signifies that as regards Mathematics SAT low levels of Reasoned dialogue relative to Non-Dialogic plus high levels of Referring Back were as productive as high levels of Reasoned dialogue relative to Non-Dialogic plus low levels of Referring Back.[7] Both were more productive than other combinations.

- Insert Table 5 and Figure 1 about here -

Relations between Elaborated, Querying and Student Participation that resemble Figure 1 were also detectable with Reading SAT, albeit not to a statistically significant degree. Thus, when the classes were subcategorized as shown in Table 6, the estimated marginal mean scores (i.e. NFER baselines included) of the group characterized as 'Above average Elaborated/Querying + Above average Student Participation' significantly exceeded those of most other groups with all three SAT scores. The following extract from one class

[7] Recall that the *lower* the ratio, the *higher* the frequency of Reasoned (or Elaborated) relative to Non-Dialogic.

that was above average along both dimensions illustrates, with Elaborated and Querying highlighted, how the combination operated in practice. The class is discussing the statement, 'It was right for the dogs to be shot', in the context of Shackleton's journey to the Antarctic:

Shackleton's Journey

Saleem: …Me and Mahir agreed that it was right, and we agreed **and also disagreed** with the statement, because they could have used the dogs for guard dogs or something like transportation.

Teacher: Okay, okay, thank you for your point. **Who would like to add or build on what Saleem's just said?**

Malika: Building on what Saleem's said, **I disagree** because they have guard dogs… say for example some of the men they were hunting and polar bears came, **then guard dogs wouldn't be enough to guard their belongings**.

Teacher: [After 4 turns] **Who would like to build on from what Malika's just said?**

Ayesha: I agree with Malika. The main priority was that it was right for them to shoot. If they hadn't shot the dogs, **the dogs will die of starvation and that's more painful than dying of a shot because, if you die of a shot, it's only painful for like one second**.

- Insert Table 6 about here -

Two-level models revealed no statistically significant main effects of dialogue with either the science test or the reasoning test, and no significant interactions relating to science. With reasoning scores, however, the Elaborated x Refer Back interaction was statistically significant and the Elaborated/Non-Dialogic Ratio x Refer Back interaction approached significance (see Table 5). The absence of significant main effects notwithstanding, reasoning

scores tended to be higher when the frequency of Elaborated was low both in absolute terms and relative to Non-Dialogic contributions, and in both cases the effect of Referring Back was to ameliorate these differences. Strangely though, it seems from Figure 2 as if the *higher* the frequency of Referring Back the less marked the impact of absolute levels of Elaborated dialogue on reasoning scores, and the *lower* the frequency of Referring Back the less marked the impact of levels relative to Non-Dialogic. Faced with inconsistency, it may be unwise to place much emphasis on either result, especially when as is clear from Figure 2's vertical axes, neither Elaborated nor Referring Back exerted marked influences in any case.

- Insert Figure 2 about here -

With end-of-year PASS, two-level modeling did, for the first time, reveal statistically significant main effects: there were positive relations between PASS scores and the frequency of Elaborated, the Elaborated/Non-Dialogic Ratio, and the Reasoned/Non-Dialogic Ratio (see Table 5). However, the latter two results were qualified through significant interactions with Querying, while the Querying x Reasoned interaction also came close to significance even though there was no significant main effect of Reasoned on PASS scores. These interactions are depicted in Figure 3, where it is clear that high levels of Elaborated or Reasoned relative to Non-Dialogic and high absolute levels of Reasoned were only productive when levels of Querying were low. When levels of Querying were high, these three variables made little difference. By contrast, the main effect of Elaborated on PASS scores was not compromised through interaction with Querying, $t(71.07)=-1.07$, $p=.288$.

- Insert Figure 3 about here -

Discussion

The study reported here was predicated upon five themes that, together with a further theme that cuts across all five, embody widely held conceptions of productive classroom

dialogue. Research into small-group interaction amongst students has endorsed some of the themes within that specific context. However, what is productive amongst students is not necessarily productive when teachers are involved, and previously there was little convincing evidence relating to teacher-student dialogue. This is despite its predominance in classroom settings. As a contribution towards filling the gap, the study sampled the teacher-student dialogue that occurs naturally in a large and demographically heterogeneous set of classrooms, analyzed this dialogue via turn-level codes and lesson-level ratings that reflected the themes, and examined how dialogue variables derived from codes and ratings related to the performance of individual students on end-of-year attainment and attitudinal scales. A thorough approach was taken towards assessing potentially confounding variables, with such variables considered as appropriate in the analyses. In the event, two forms of dialogue were identified whose frequencies, so long as students participated extensively, were positively associated with SAT scores: Elaborated, derived from ELI and EL in Table 1, and Querying, as defined in Table 1. When student participation was limited the frequencies of these variables were irrelevant, just as extensive participation had little impact unless the frequencies of Elaborated and Querying were high. The frequency of Elaborated dialogue was also positively associated with PASS scores, and unlike with other similarly associated variables the relation was not undermined through high frequency Querying.

These results are consistent with the first of the underlying themes, i.e. that initiations should include open questions. Invitations to elaborate (ELI) are included in the Elaborated variable, and as with 'Who would like to build on from what Malika's just said?' such invitations are inherently open. Via the inclusion of Elaborated, the results are also self-evidently consistent with the second theme, i.e. that participants should make extended contributions, elaborating and building on previous contributions from themselves and others. The involvement of Querying resonates with the third theme's stipulation that differences be

acknowledged and probed, and when the positive effects of Elaborated and Querying depend upon high levels of student participation, the results also endorse the crosscutting theme, i.e. that all participants should contribute fully, not merely teachers. However, while Elaborated or Querying plus Student Participation have emerged as positive predictors and as such offer support for the themes, there are grounds for circumspection. Elaborated, Querying and/or Student Participation were not positively associated with performance on the science and reasoning tests. Indeed, with reasoning, there was a somewhat negative association between Elaborated dialogue and performance, albeit one that was qualified through Referring Back. Furthermore, Elaborated, Querying and Student Participation were merely three dialogue variables in a set of 11, with the remaining eight also indicative of theoretically productive classroom dialogue. Little supportive evidence was obtained for any of these eight, and when one of them (Reasoned) bears on the first theme (by virtue of including REI, which are open questions) and the third theme (by virtue of covering the reasons on which opinions are based), all-round endorsement of these two themes has not been obtained. This is despite the support for the themes offered through Elaborated, Querying and Student Participation.

Elaborated and Querying with Student Participation

Faced with this picture, fuller assessment of the themes' significance seems to require clarity about three major issues, namely why Elaborated, Querying and Student Participation emerged as positive predictors with some outcome variables, why their value did not extend to the science and reasoning tests, and why other dialogue variables (especially, perhaps, those relating to reasoned dialogue) did not prove productive. To address the first issue, the Student Participation scale needs to be examined more deeply. So far, its specification has been limited to the fact that its three points represent different levels of student input. The points were in fact associated with detailed descriptors: 0=*Public exchanges consist in questioning and succinct students' contributions or students don't have opportunities to*

discuss their ideas publicly; 1=*Students express their ideas publicly at length in whole-class situation and group work, but they don't engage with each other's ideas*; 2=*Multiple students express their ideas publicly at length in whole-class situation and group work and in doing so, they engage with each other's ideas, for example by referring back to their contributions, challenging or elaborating on them. This includes spontaneous or teacher-prompted participation*. Thus, when during the *Shackleton's Journey* extract quoted above Malika and Ayesha not only expressed ideas but also responded to the preceding student's ideas, they were exhibiting behavior consistent with a rating of 2.

Engagement with other students' ideas does not necessitate elaborating or querying, just as those two forms of dialogue are possible without referring to the ideas of others. However, the teacher's interventions in the *Shackleton's Journey* extract more-or-less guaranteed elaboration in that particular context and, given probable divergence of opinion, made querying very likely. Because the two forms were in fact triggered and student contribution was also high, the upshot, as shown, was a rich tapestry of inter-connected and discursively salient ideas. Moreover, it was a tapestry in which each student's individual ideas had been made explicit together with plausible alternatives, and this could be significant. Within the literature, the juxtaposition in dialogue of own ideas with those of others has been depicted repeatedly as initiating reflective comparison and appraisal, and through this, insight and learning. Amongst classic theorists, Piaget (1959, p.137) saw value in juxtaposition because it allows each individual to see 'himself in the eyes of others and thus acquire the habit of watching himself think'. For Vygotsky (1998, p.168), it plays an indispensable role in confronting the individual with 'the need to form a basis, to prove, confirm and verify his own idea'. In Bakhtin (1981, p.348), struggling with another's discourse was regarded as the key to 'ideological consciousness', and through this, greater understanding within those who struggle. Construed in contemporary terms, these theorists

were arguing that juxtaposition in dialogue engenders a 'meta-cognitive' perspective upon personal beliefs, and this is the source of its value. Certainly, there is ample evidence that adopting a meta-cognitive perspective supports knowledge growth (Hattie, 2009; Higgins, 2013). Perhaps then, this is why high levels of Elaborated and Querying together with high Student Participation proved productive. They created conditions where students could 'watch themselves think', and it was this that triggered growth.

Science and Reasoning

Yet the benefits from Elaborated, Querying and Student Participation did not extend to the science and reasoning tests, and to understand fully how those dialogue variables operate it seems important to establish the reasons behind this lack of relationship. The primary issue appears to be whether progress here is immune to the key forms of dialogue, or whether extraneous factors might in this particular study have masked their potential significance. The former would imply that there are types of attainment that lie beyond the scope of Elaborated, Querying and Student Participation; the latter would mean that lying within the scope depends on additional factors, which the above meta-cognitively oriented model would need to acknowledge. As regards extraneous factors, one possible source lies with the tests themselves, for the science and reasoning tests were the only outcome measures to be designed specifically for the study and therefore not tried-and-tested. However, the tests have both emerged as psychometrically robust, and in terms of content, the science test maps onto the prescribed Year 6 curriculum. As regards test processing, there is no reason to question the largely teacher-presented delivery: SATs and (usually) PASS were also teacher presented. Moreover, test marking was checked, and found to be reliable. Yet while test-related problems do not seem plausible, there were, with science, further factors extraneous to dialogue that might have been operating. One such factor lies with sampling, for the science test was taken in only three of the classes depicted in Table 6 as above average with

both Elaborated/Querying and Student Participation. This reflects proportionally higher attrition than expected when 61% of the classes covered inheritance and evolution and therefore used the test, and also than was associated with the table's other cells. It could be a chance phenomenon, but equally it could be strategic. Perhaps teachers who measured well on the key dialogue variables omitted inheritance and evolution because they perceived the topic to be incompatible with their teaching styles. Beyond this, however, compromising factors were signaled even when the topic was covered. It was often left until after SATs, i.e. close to the year's end; some teachers reported coverage at lower levels than the curriculum prescribes; and no doubt reflecting this the students found the test difficult, with a sample mean score of only 13.99 out of a possible 33. If such trends are true of science in general and not merely the specific topic (and the test's procedural component was generic), they most likely reflect the marginalization that removing primary science SATs has triggered (Leonard, Lamb, Howe, & Choudhoury, 2017). In any event, limited motivation to master the subject matter is a further factor extraneous to dialogue that, in the context of science, may have moderated the impact of Elaborated, Querying and Student Participation.

With sampling and motivational limitations possible, it would be premature to suggest that Elaborated, Querying and Student Participation have no potential relevance in the context of science. Further research is needed. For that reason, it seems unnecessary at this stage to be concerned at apparent discrepancies with the work of Muhonen et al. (2018), which reports high ratings for 'building upon' during dialogue being positively associated with attainment in science. Yet while a non-committal stance as regards Elaborated, Querying and Student Participation seems appropriate with the science test, adopting this stance with the reasoning test looks harder to justify. Here, there was no sample attrition for the students in all 72 classes took the reasoning test, and with a sample mean score of 24.01 out of a possible 40 the extreme challenge of the science test was not repeated. Moreover, the

preliminary analyses indicated a positive relation between the quality of student group work and reasoning scores, suggesting that the students took the test seriously. The quality of group work was of course assessed against features of dialogue that previous research has shown to be productive in that context, indicating that dialogue probably contributed to the positive relation with test scores. However, the relevant features amongst the group work scales revolve around reasoned discussion rather than anything straightforwardly relatable to Elaborated and Querying. Furthermore, the picture that Figure 2 paints for teacher-student interaction should not be forgotten: while the values did not reach conventional levels of statistical significance, the relation between scores on the reasoning test and Elaborated dialogue (and the Elaborated/Non-Dialogic ratio) tended to be negative. All in all then, it appears that the skills tapped via the reasoning test might genuinely be immune to the influences of Elaborated, Querying and Student Participation, implying limits on the scope of these otherwise productive variables. If dialogue is relevant for reasoning skills, it looks instead to be reasoned dialogue of the type that occurs during group work amongst students.

Reasoned Dialogue

Yet while the relation between quality of student group work and reasoning test scores suggest a role for reasoned dialogue in the small-group context, there was no evidence for a contribution from such dialogue when it occurred during interactions involving teachers. Neither the absolute frequency of the study's Reasoned variable, nor its frequency relative to Non-Dialogic, was associated with scores on the reasoning test. Indeed, neither variable was positively related to SAT scores or to scores on the science test. Across these tests, there was only one statistically significant result involving the variables that addressed reasoned dialogue, namely the Reasoned/Non-Dialogic Ratio x Refer Back interaction detected with Mathematics SAT. However, this did not reflect a positive impact of ratio. Rather, ratio operated differently depending on the frequency of Referring Back but had no net benefits.

With end-of-year PASS, the Reasoned/Non-Dialogic Ratio was positively associated with scores (as to a non-significant extent was the Reasoned variable), but the benefits in both cases were eliminated in contexts of high frequency Querying. Since high frequency Querying is desirable given its positive consequences elsewhere, even the results with PASS cannot be seen as encouraging.

The absence of positive relationships around reasoned dialogue was not an artifact of the variables' distributional properties. As Table 2 indicates, both the Reasoned variable and the Reasoned/Non-Dialogic Ratio were associated with adequate mean frequencies and wide ranges across classrooms. Remembering Webb et al.'s (2009) distinction between high and low quality reasons, it is possible that the results would have been more encouraging given a more nuanced conception of reasons. However, elaborations could in principle also be divided into high and low quality, yet the Elaborated variable proved positively predictive without division. Of greater relevance perhaps is the point made above in relation to the reasoning test, that performance may have been associated with reasoned dialogue *of the type that occurs during group work amongst students*. During group work, reasons are typically expressed to justify each participant's proposals when opinions differ, e.g. 'Let's say the ball will float', 'No, let's write sink because it's really heavy', and 'But ships float and they're heavy; the ball's hollow like ships and that's why it'll float' (Howe, 2009, 2010; Howe & Zachariou, 2017). Thus, reasons characteristically occur in what earlier was referred to as 'a rich tapestry of inter-connected and discursively salient ideas' about whose status students are motivated to reflect. This may explain their productiveness. By contrast, when reasons occurred in the present study's teacher-student dialogues, they were often solicited by teachers to check and provide feedback on understanding. Thus, even when multiple reasons were expressed, as with the discussion below about which rational number terms should be

inserted into crossword boxes, there was no uncertainty about their quality and therefore no need for contrastive reflection from students.

Rational Number Crossword

Teacher: Okay, who is feeling brave then? Who's willing to make a mistake for the rest of the class? Or who thinks they can justify their answer and maybe they're correct - who knows? Go on Nick.

Nick: Hundredth.

Teacher: Hundredth? Okay and your justification please?

Nick: Because it fits in the boxes.

Teacher: Nick! Do we have a different justification please? Nick, listen carefully young man. Carl please?

Carl: Cos it says 'this place value'.

Teacher: Okay. That's a good place to start Carl, that's a great place to start - we know we're after the name of a place value.

Of course, had the teacher drawn attention to the difference between Nick and Carl's reasons and invited reflection on their merits, the putative conditions for progress would have been created. However, that would have required what here is termed Reasoned Co-ordination, and as Table 1 shows RC moves were extremely rare. As a result, the present study is limited to the Reasoned and Reasoned/Non-Dialogic Ratio variables, and the conclusion as regards these variables is that they had no straightforwardly positive consequences. The motivation for including the variables was, of course, the third theme associated with theoretically productive classroom dialogue, i.e. that differences of opinion should not merely be acknowledged and probed as reflected in Querying, but also examined with reference to the reasons on which opinions are based. The theme features prominently in the literature, yet as regards teacher-student dialogue there is little compelling empirical

research. Even the studies that spotlight the issue (e.g. O'Connor et al., 2015; Pauli & Reusser, 2015; van der Veen et al., 2017) involve data that preclude isolating the consequences of reasoned dialogue from the effects of other forms. Thus, the mixed results relating to the third theme (encouraging with Querying, but not with Reasoned or the Reasoned/Non-Dialogic Ratio) may perhaps be less surprising than they might initially seem.

Limitations and Conclusions

The fourth theme associated with theoretically productive classroom dialogue was that through explicit links amongst contributions and attempts to co-ordinate, integrated lines of enquiry should be pursued. The results obtained with Referring Back and Referring Widely bear on this theme, and with Referring Back the data in Figure 2 signal relevance for student outcome. Yet it is hard to discern a consistent pattern to these data, and as regards Referring Widely there were no significant main or interaction effects in the first place. So insofar as the fourth theme could be explored, the results are not especially encouraging. Equally though, the exploration that could take place should not be treated as conclusive. The frequencies of Referring Back and Referring Widely were low relative to other variables (see Table 1), and they may have been insufficiently prominent to reveal their potential (see also Howe et al., 2007). In addition, the frequencies of the other variables relevant to the fourth theme, i.e. the three indices of co-ordination (CI, SC, and RC in Table 1), were so low that analysis proved impossible. Regrettably, equivalent problems beset the fifth theme relating to the adoption of a meta-cognitive perspective. While the preceding discussion has underlined the potential significance of this perspective, the value of dialogue promoting its adoption could not be adequately examined. It is true that the Monitoring and Refection scales produced usable ratings (which were never associated with statistically significant effects), and these scales were regarded as addressing meta-cognition. However, it was the Talk Rules

scale that was seen as the major index and ratings using this scale were too consistently low (and potentially unreliable) to be taken further.

It was anticipated from the outset that the naturalistic methodology would be subject to such limitations, but it was also recognized that an interventionist approach would not necessarily prove more successful. For instance, an intervention that aspired amongst other things to boost co-ordination (Ruthven et al., 2017) resulted in frequencies for the relevant variables that are very similar to those shown in Table 1 for CI, SC and RC. Yet the uncertainties remaining after the present study need to be addressed, and the need is not restricted to the fourth and fifth themes. The intricate inter-relations amongst those dialogue variables that could be examined require further analysis, as do the broader contextual factors that may constrain their operation. Such factors may include those aspects of classroom ethos encompassed, as noted earlier, within the concept of dialogic pedagogy, but remembering the science test they may equally include participant motivation. Cultural differences might also be considered for although the study's sample was demographically diverse, it was nevertheless restricted to England. Perhaps the best way forward would be to plan highly targeted interventions, for these might pre-empt the problems associated with broader approaches. For instance, levels of Simple Co-ordination (and only this) could be promoted in classrooms where Elaborated, Querying and Student Participation are already embedded in teacher-student dialogue to ascertain whether this adds value as regards SATs and PASS. Reasoned Co-ordination could be boosted during teacher-student interactions that are already high in reasoned dialogue, with a view to observing whether Reasoned and/or the Reasoned Non-Dialogic Ratio are then related to scores on a reasoning test (and perhaps also whether the benefits exceed those obtainable from student group work).

Viewed in this light, the results of the present study should be regarded as a staging post in a lengthy and iterative research process. Many future studies will be required before

the optimal patterns of classroom dialogue can be specified in full. Nevertheless, the results make an original and significant contribution in their own right: this is the first large-scale investigation of the relationship between teacher-student dialogue and student outcomes which systematically takes account of relevant confounds. As regards the concept of productive classroom dialogue that guided the study, the results seem compatible with several conclusions. As noted, they offer partial but not whole-hearted support for both the first theme (when some but not all types of open question have emerged as productive) and the third theme (when value has been detected in acknowledging and probing differences of opinion, but not, at least as currently effected in teacher-student dialogue, in examining the reasons on which opinions are based). The repeated relevance of the Elaborated variable implies stronger support for the second theme, yet even here the results with the reasoning test highlight exceptions. Faced with this picture, it might be concluded that the themes (and therefore the concept of productive dialogue) are overly general, and it would be preferable in the future to focus upon more specific constructs, perhaps along the lines of the study's variables. However, as emphasized already, the themes have dominated the research literature, suggesting that it may be premature to lose sight of them. It may be best, for now at least, to treat the themes as frameworks within which sub-types can be identified. The consequences of these sub-types may (and, from the present results, do) turn out to differ.

In addition to their prominence within the research literature, the themes have also been used to inform teacher professional development with a view to changing classroom practice. While the study was intended to clarify the themes rather than develop their practical implications, its results do carry important implications for teaching. Teacher-student dialogue that manifests high levels of Elaborated, Querying, and Student Participation goes beyond the simple I-R-E/F format with which the paper began, but as the *Shackleton's Journey* extract illustrates, these variables are readily superimposed upon that structure. Is

and Rs need to be embellished and teacher E/Fs need to be withheld to permit student reflection and appraisal, but the basic format can be preserved. Indeed, given the prevalence of OI and UC turns (and therefore Non-Dialogic) documented in Tables 1 and 2, the key variables almost certainly *are* superimposed on I-R-E/F in classrooms where they are already used with high frequency. The implication is that Elaborated, Querying, and Student Participation comprise a relatively straightforward combination of dialogue features that, referring to exemplary instances like *Shackleton's Journey*, could be promoted to teachers and manageably achieved in practice. Achievement of this combination alone should pay dividends. Whether the dividends will be optimal remains to be seen: the need to resolve the uncertainties and context dependencies highlighted above bears re-iteration. Nevertheless, benefits should accrue from promoting Elaborated, Querying, and Student Participation and recognizing this, members of the team responsible for the study have begun working on professional development programmes and resources that will support classroom application (see http://bit.ly/T-SEDA). The work includes detailed specification of the strategies teachers can use to draw out Elaborated dialogue and Querying in highly participative classrooms.

References

Adey, P., & Shayer, M. (2015). The effects of cognitive acceleration. In L.B. Resnick, C.S.C. Asterhan, & S.N. Clarke (Eds.), *Socializing intelligence through academic talk and dialogue* (pp.127-140). Washington, D.C: American Educational Research Association.

Ahmed, A., Howe, C., Major, L., Hennessy, S., Mercer, N., & Warwick, P. (Forthcoming). Developing a test of reasoning for preadolescents. Manuscript under review.

Alexander, R. (2008). *Towards dialogic teaching: Rethinking classroom talk*. Cambridge: Dialogos.

Alexander, R., Hardman, F., & Hardman, J. (2017). *Changing talk, changing thinking.* Interim Report from the In-house Evaluation of the CPRT/UoY Dialogic Teaching Project. Retrieved from www.robinalexander.org.uk

Anderson, A., Howe, C., Soden, R., Halliday, J., & Low, J. (2001). Peer interaction and the learning of critical thinking skills in further education students. *Instructional Science, 29,* 1-32. doi:10.1023/A:1026471702353

Applebee, A.N., Langer, J.A., Nystrand, M., & Gamoran, A. (2003). Discussion-based approaches to developing understanding: Classroom instruction and student performance in middle and high school English. *American Educational Research Journal, 40,* 685-730. doi:10.3102/00028312040003685

Bakhtin, M. (1981). *The dialogic imagination: Four essays*. Austin: University of Texas Press.

Benjamini, Y., Krieger, A.M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika, 93,* 491-507. doi:10.1093/biomet/93.3.491

Boyd, M.P., & Markarian, W.C. (2011). Dialogic teaching: Talk in service of a dialogic stance. *Language and Education, 25,* 515-534. doi:10.1080/09500782.2011.597861

Chin, C. (2007). Teacher questioning in science classrooms: Approaches that stimulate productive thinking. *Journal of Research in Science Teaching, 44,* 815-843. doi:10.1002/tea.20171

Chinn, C.A., & Anderson, R.C. (1998). The structure of discussions that promote reasoning. *Teachers College Record, 100*(2), 315-368.

Chinn, C.A., Anderson, R.C., & Waggoner, M.A. (2001). Patterns of discourse in two kinds of literature discussion. *Reading Research Quarterly, 36,* 378-411. doi:10.1598/rrq.36.4.3

Clarke, D., Xu, L.H., & Wan, M.E.V. (2010). Student speech as an instructional priority: Mathematics classrooms in seven culturally-differentiated cities. *Procedia Social and Behavioral Sciences*, *2,* 3811-3817. doi:10.1016/j.sbspro.2010.03.595

Damon, W., & Phelps, E. (1989). Critical distinctions among three approaches to peer education. *International Journal of Educational Research, 5,* 9-19. doi:10.1016/0883-0355(89)90013-X

Daniels, H. (2001). *Vygotsky and pedagogy*. Abingdon: Routledge.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching.* Alexandria, VA: Association for Supervision and Curriculum Development.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics.* Sage: Los Angeles.

Firestone, G., & Brody, N. (1975). Longitudinal investigation of teacher-student interactions and their relations to academic performance. *Journal of Educational Psychology. 67,* 544-550. doi:10.1037/h0076994

Freire, P., & Macedo, D. (1995). A dialogue: Culture, language and race. *Harvard Educational Review, 65,* 377-402. doi:10.17763/haer.65.3.12g1923330p1xhj8

Fung, D., & Howe, C. (2014). Group work and the learning of critical thinking in the Hong Kong secondary liberal studies curriculum. *Cambridge Journal of Education, 44,* 245-270. doi:10.1080/0305764X.2014.897685

Ginsburg, A., & Smith, M.S. (2016). *Do randomized controlled trials meet the "gold standard"? A study of the usefulness of RCTs in what works clearinghouse.* Report from American Enterprise Institute. Retrieved from www.carnegiefoundation.org

GL Assessment. (2013). *PASS: Pupil attitudes to self and school.* London: GL Assessment.

Haneda, M., & Wells, G. (2008). Learning an additional language through dialogic inquiry. *Language and Education, 22*, 114-136. doi:10.2167/le730.0

Haneda, M., Teemant, A. & Sherman, B. (2017). Instructional coaching through dialogic interaction: Helping a teacher to become agentive in her practice. *Language and Education, 31,* 46-64. doi:10.1060/09500782.2016.1230127

Hattie, J.A. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* London: Routledge.

Hennessy, S., Dragovic, T., & Warwick, P. (2017). A research-informed, school-based professional development workshop programme to promote dialogic teaching with interactive technologies. *Professional Development in Education, 44,* 1-24. doi:10.1080/19415257.2016.128653

Hennessy, S., Rojas-Drummond, S., Higham, R., Márquez, A.M., Maine, F., Ríos, R.M., … Barrera, M.J. (2016). Developing a coding scheme for analysing classroom dialogue across educational contexts. *Learning, Culture and Social Interaction, 9,* 16-44. doi:10.1016/j.lcsi.2015.1.2.001

Herrenkohl, L.R., Palincsar, A.S., DeWater, L.S., & Kawasaki, K. (1999). Developing scientific communities in classrooms: A sociocognitive approach. *The Journal of the Learning Sciences, 8,* 451-493. doi:10.1080/10508406.1999.9672076

Higgins, S. (2013). Self-regulation and learning: Evidence from meta-analysis and from classrooms. *British Journal of Educational Psychology Monograph Series II, 10,* 111-126.

Howe, C. (2009). Collaborative group work in middle childhood: Joint construction, unresolved contradiction and the growth of knowledge. *Human Development, 52,* 215-239. doi:10.1159/000215072

Howe, C. (2010). *Peer groups and children's development.* Oxford: Blackwell.

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

Howe, C., & Mercer, N. (2007). *Children's social development, peer interaction and classroom learning.* The Primary Review (Research Survey 2/1b). Cambridge: University of Cambridge.

Howe, C., Tolmie, A., Duchak-Tanner, V., & Rattray, C. (2000). Hypothesis testing in science: Group consensus and the acquisition of conceptual and procedural knowledge. *Learning and Instruction, 10,* 361-391. doi:10.1016/S0959-4752(00)00004-9

Howe, C., Tolmie, A., Thurston, A., Topping, K., Christie, D., Livingston, K., … Donaldson, C. (2007). Group work in elementary science: Towards organisational principles for supporting pupil learning. *Learning and Instruction, 17,* 549-563. doi:10.1016/j.learninstruc.2007.09004

Howe, C., & Zachariou, A. (2017). Small-group collaboration and individual knowledge acquisition: The processes of growth during adolescence and early adulthood. *Learning and Instruction.* Advance online publication. doi:10.1016j.learninstruc.2017.10.007

Hughes, D.C. (1973). An experimental investigation of the effects of pupil responding and teacher reacting on pupil achievement. *American Educational Research Journal, 10,* 21-37. doi:10.3102/00028312010001021

Jurkowski, S., & Hänze, M. (2015). How to increase the benefits of cooperation: Effects of training in transactive communication on cooperative learning. *British Journal of Educational Psychology, 85,* 357-371. doi:10.1111/bjep.12077

Kumpulainen, K., & Lipponen, L. (2010). Productive interaction as agentic participation in dialogic enquiry. In K. Littleton & C. Howe (Eds.), *Educational dialogues: Understanding and promoting productive interaction* (pp.48-63). London: Routledge.

Kutnick, P., & Blatchford, P. (2014). *Effective group work in primary school classrooms: The SPRinG approach.* Dordrecht: Springer.

Larrain, A., Howe, C., & Friere, P. (2018). 'More is not necessarily better': Curriculum materials support the impact of classroom argumentative dialogue in science teaching on content knowledge. *Research in Science and Technological Education, 36,* 282-301.. doi:10.1080/02635143.2017.1408581

Lefstein, A. (2010). More helpful as problem than solution: Some implications of situating dialogue in classrooms. In K. Littleton & C. Howe (Eds.), *Educational dialogues: Understanding and promoting productive interaction* (pp.170-191). London: Routledge.

Lefstein, A., & Snell, J. (2014). *Better than best practice: Developing teaching and learning through dialogue.* London: Routledge.

Leonard, S., Lamb, H., Howe, P., & Choudhoury, A. (2017). *'State of the nation' report of UK primary science education.* Baseline Research for the Wellcome Trust Primary Science Campaign. Retrieved from https://wellcome.ac.uk/sites/default/files/state-of-the-nation-report-of-uk-science-education.pdf

Littleton, K., & Howe, C. (2010). Introduction. In K. Littleton & C. Howe (Eds.), *Educational dialogues: Understanding and promoting productive interaction* (pp.1-7). London: Routledge.

Luckner, A.E., & Pianta, R.C. (2011). Teacher-student interactions in fifth grade classrooms: Relations with children's peer behavior. *Journal of Applied Developmental Psychology, 32,* 257-266. doi:10.1016/j.appdev.2011.02.010

Maas, C.J.M., & Hox, J.J. (2005). Sufficient sample size for multilevel modeling. *Methodology, 1,* 86-92. doi:10/1027/1614-1881.1.386

Matusov, E. (2009). *Journey into dialogic pedagogy.* New York: Nova Science.

Mehan, H. (1979). *Learning lessons: Social organization in the classroom.* Cambridge, MA: Harvard University Press.

Mercer, N., & Dawes, L. (2008). The value of exploratory talk. In N. Mercer & S. Hodgkinson (Eds.), *Exploring talk in school* (pp.55-71). London: Sage.

Mercer, N., Dawes, R., Wegerif, R., & Sams, C. (2004). Reasoning as a scientist: Ways of helping children to use language to learn science. *British Educational Research Journal*, *30,* 367-385. doi:10.1080/01411920410001689689

Mercer, N., & Littleton, K. (2007). *Dialogue and the development of children's thinking: A sociocultural approach.* London: Routledge.

Mercer, N., & Sams, C. (2006). Teaching children how to use language to solve maths problems. *Language and Education*, *20,* 507-528. doi:10.2167/le678.0

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

Miller, G.A., & Chapman, J.P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology, 110,* 40-48. doi:10.1037/0021-843X.110.1.40

Mortimer, E.F., & Scott, P.H. (2003). *Meaning making in science classrooms.* Buckingham: Open University Press.

Muhonen, H., Pakarinen, E., Poikkeus, A-M., Lerkkanen, M-K., & Rasku-Puttonen, H. (2018). Quality of educational dialogue and association with students' academic performance. *Learning and Instruction, 55,* 67-79. doi:10.1016/j.learninstruc.2017.09.007

Nystrand, M. (2006). Research on the role of classroom discourse as it affects reading comprehension. *Research in the Teaching of English, 40*(4)*,* 392-412.

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,* 135-198. doi:10.1207/S15326950DP3502_3

O'Connor, C., Michaels, S., & Chapin, S. (2015). 'Scaling down' to explore the role of talk in learning: From district intervention to controlled classroom study. In L.B. Resnick, C.S.C. Asterhan, & S.N. Clarke (Eds.), *Socializing intelligence through academic talk and dialogue* (pp.111-126)*.* Washington, D.C: American Educational Research Association.

Osborne, J., Erduran, S., Simon, S., & Monk, M. (2001). Enhancing the quality of argument in school science. *School Science Review, 82 (301),* 63-70.

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,* 315-347. doi: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, D.C: American Educational Research Association.

Pehmer, A.K., Gröschner, A., & Seidel, T. (2015). Fostering and scaffolding student engagement in productive classroom discourse: Teachers' practice changes and reflections in light of teacher professional development. *Learning, Culture and Social Interaction*, *7,* 12-27. doi:10.1016/j.lcsi.2015.05.001

Piaget, J. (1959). *Judgment and reasoning in the child.* Michigan: Littlefield Adams.

Pimentel, S.D., & McNeill, K.L. (2013). Conducting talk in secondary science classrooms: Investigating instructional moves and teachers' beliefs. *Science Education, 97,* 367-394. doi:10.1002/sce.21061

Preacher, K.J., & Hayes, A.F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments & Computers, 36,* 717-731. doi:10.3758/BF03206553

Reznitskaya, A., & Gregory, M. (2013). Student thought and classroom language: Examining the mechanisms of change in dialogic teaching. *Educational Psychologist, 48,* 114-133. doi:10.1080/00461520.2013.775898

Reznitskaya, A., Kuo, L-J., Clark, A-M., Miller, B., Jadallah, M., Anderson, R.C., & Nguyen-Jahiel, K. (2009). Collaborative reasoning: A dialogic approach to group discussions. *Cambridge Journal of Education, 39,* 29-48. doi:10.1080/03057640802701952

Richter, F.D., & Tjosvold, D. (1980). Effects of student participation in classroom decision making on attitudes, peer interaction, motivation, and learning. *Journal of Applied Psychology, 65,* 74-80. doi:10.1037/0021-9010.65.1.74

Rojas-Drummond, S., Littleton, K., Hernandez, F., & Zuniga, M. (2010). Dialogical interactions among peers in collaborative writing contexts. In K. Littleton & C. Howe

(Eds.), *Educational dialogues: Understanding and promoting productive interaction* (pp.128-148). London: Routledge.

Ruthven, K., Mercer, N., Taber, K.S., Guardia, P., Hofmann, R., Ilie, S., … Riga, F. (2017). A research-informed dialogic-teaching approach to early secondary school mathematics and science: The pedagogical design and field trial of the epiSTEMe intervention. *Research Papers in Education, 32,* 18-40. doi:10.1080/02671522.2015.1129642

Schwarz, B.B. (2009). Argumentation and learning. In N.M. Mirza & A-N. Perret-Clermont (Eds.), *Argumentation and education: Theoretical foundations and practices* (pp.91-126). Dordrecht: Springer.

Sedova, K., Sedlacek, M., & Svaricek, R. (2016). Teacher professional development as a means of transforming student classroom talk. *Teaching and Teacher Education, 57,* 14-25. doi:10.1016/j.tate.2016.03.005

Sinclair, J.Mc.H., & Coulthard, M. (1975). *Towards an analysis of discourse: The English used by pupils and teachers.* Oxford: Oxford University Press.

Trickey, S., & Topping, K. (2004). Philosophy for children: A systematic review. *Research Papers in Education, 19,* 365-380. doi:10.1080/0267152042000248016

van der Veen, C., de Mey, L, van Kruistum, C., & van Oers, B. (2017). The effect of productive classroom talk and metacommunication on young children's oral communicative competence and subject matter knowledge: An intervention study in early childhood education. *Learning and Instruction, 48,* 14-22. doi:10.1016/j.learninstruc.2016.06.001

Vrikki, M., Wheatley, L., Howe, C., Hennessy, S., & Mercer, N. (Forthcoming). Dialogic practices in primary school classrooms. In press at *Language and Education.*

Vygotsky, L. (1998). *The collected works of L. S. Vygotsky. Volume 5: Child psychology.* New York: Plenum Press.

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,* 49-70. doi:10.1080/03057640802701986

Wegerif, R. (2013). *Dialogic: Education for the internet age.* London: Routledge.

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

Wertsch, J.V. (1990). The voice of rationality in a sociocultural approach to mind. In L.C. Moll (Ed.), *Vygotsky and education: Instructional implications and applications of sociohistorical psychology* (pp.111-126). Cambridge: Cambridge University Press.

Wilkinson, I.A.G., Reznitskaya, A., Bourdage, K., Oyler, J., Glina, M., Drewry, R., … Nelson, K. (2017). Toward a more dialogic pedagogy: Changing teachers' beliefs and practices through professional development in language arts classrooms. *Language and Education, 31,* 65-82. doi:10.1080/09500782.2016.1230129

Appendix: Identification of Confounds

Student Characteristics					
Potential confound	**Assessment procedures**	**Variable**	**Sample variation**	**Dialogue related**	**Outcome related**
Prior attainment	NFER: Maths (Test 1 – written component, max. score = 35) Variables: Class mean and SD	PreMaths (M)	√	√	√
		PreMaths (SD)	√	√	√
	NFER: Reading (max. score = 44) Variables: Class mean and SD	PreRead (M)	√	√	√
		PreRead (SD)	√	X	na
Engagement Motivation Attitudes Self-concept	CQ: PASS. 50-item questionnaire, e.g. *I am bored at school, I enjoy doing hard schoolwork, I think this is a good school, I am clever*. Response options (scores): No, not at all (1), No, not much (2), Yes, a bit (3), Yes, a lot (4). Variable: Class mean (excluding students omitting >5 items)	PrePASS	√	√	√
Parental involvement	CQ: 5 items about discussion e.g. *How often do you talk to your parent or guardian about things you have learned in class?* Response options (scores): Less than once a week (1), Once or twice a week (2), Most days (3). 4 items about visits, e.g. *In the last year has your parent or guardian spoken to your teacher?* Response options (scores): No (1), Don't know (2), Yes (3). Variables: Class means (excluding students omitting >1 item)	PI_Talk	√	X	na
		PI_Visit	√	X	na
Mobility	CQ: *How many schools have you been to before coming to this one?*	Mobility	√	X	na

	Variable: Number provided				
Socioeconomic status	TQ: *How many children are in your class? How many of them are eligible for free school meals?* Variable: Percentage ineligible	%NoFSM	√	X	na
English as second language	TQ: *For how many children is English not the first language?* Variable: Percentage whose first language is English	%NoESL	√	X	na
Fluency in English	TQ: *How many children are fluent in English?* Variable: Percentage fluent	%Fluent	X	na	na
Special needs	TQ: *How many children are registered with special needs?* Variable: Percentage without special needs	%NoSEN	√	X	na

Teacher Practices

Potential confound	**Assessment procedures**	**Variable**	**Sample variation**	**Dialogue related**	**Outcome related**
Homework	TQ: *How often do the children in your class do homework?* Response options (scores): Never (1), Once or twice a year (2), Once or twice a term (3), Several times a term (4), At least once a week (5). Variable: Score for selected option	Homework	X	na	na
External trips	TQ: *How often do the children in your class go on out-of-school trips?* Response options, scores, and variable as for Homework	Trips	√	√	X
Physical exercise	TQ: *How often do the children in your class take physical exercise?* Response options, scores, and variable as for Homework	Exercise	X	na	na
Written tests	TQ: *How often do the children in your class take written tests?* Response options, scores, and variable as for Homework	Tests	√	X	na

Calculator use	TQ: *How often do the children use calculators in maths?* Response options, scores, and variable as for Homework	Calculators	√	X	na
Computer use	TQ: *How often do the children use computers in class?* Response options, scores, and variable as for Homework	Computers	X	na	na
IWB use	TQ: *How often do the children use the interactive whiteboard or interactive display screen?* Response options, scores, and variable as for Homework	IWB	X	na	na
Videos or animations	TQ: *How often do the children use video, simulation/animation or other visual media?* Response options, scores, and variable as for Homework	Video/ Anim	√	X	na
Concept mapping	TQ: *How often do the children use concept mapping or mind mapping?* Response options, scores, and variable as for Homework	Mapping	√	X	na
Thinking skills	TQ: *In your class, how often do you teach generic thinking skills?* Response options, scores, and variable as for Homework	Thinking	√	X	na
Integrated teaching	TQ: *In your class, how often do you integrate two or more areas of the curriculum?* Response options, scores, and variable as for Homework	Integration	√	X	na
Lecture	TQ: *I present new topics to the class through lecture-style presentation*. Response options: Strongly disagree (1), Disagree (2), Neither agree nor disagree (3), Agree (4), Strongly agree (5). Variable: Score for selected option	Lecture	√	X	na
Worked examples	TQ: *I use worked examples*. Response options, scores, and variable as for Lecture	WorkedEG	√	X	na
Spaced practice	TQ: Two items, e.g. *We revisit previously taught topics*. Response options and scores as for Lecture. Variable: Mean score across items	Spaced	√	X	na

Inductive teaching	TQ: Two items, e.g. *I start a topic with concrete examples and then teach the theory*. Response options, scores, and variable as for Spaced	Inductive	√	X	na
Collaborative group work	CO: Quality of observed group work rated on scales, e.g. *Group interaction involved justified reasoning, All pupils were involved in the interaction*. Variable: Mean rating across scales for all group sessions observed in each classroom	Group	√	√	√
Peer tutoring		PeerTutor	X	na	na
Goal setting	VR: Clarity of lesson goals rated: Poor (0), Moderate (1), Good (2) Variable: Total rating across two lessons per classroom	Goals	√	√	X
Feedback	VR: Quality of feedback rated. Response options, scores, and variable as for Goals	Feedback	√	√	X
Classroom relations	VR: Quality of teacher-student interactions rated. Response options, scores, and variable as for Goals	Relations	√	√	X
Behavior management	VR: Teachers' awareness of (and response to) student (mis)behavior rated. Response options, scores, and variable as for Goals	Behavior	√	X	na

Note: NFER=National Foundation for Educational Research; CQ=Child Questionnaire; PASS=Pupil Attitude to Self and School; TQ=Teacher Questionnaire; CO=Classroom Observation; VR=Video rating; na=Not applicable

Table 1

Turn-level Codes

Code	Definition	Mean (SD) frequency
Invite elaboration (ELI)	Invites building on, elaboration, evaluation, clarification of own or another's contribution. E.g. 'Have you noticed anything else that the poet uses?'	27.99 (14.06)
Elaboration (EL)	Builds on, elaborates, evaluates, clarifies own or other's contribution. E.g. [After 'It's sort of describing how you do it] 'Yes, it's got a good emphasis and a good use of vocabulary'	74.07 (27.08)
Invite reasoning (REI)	Explicitly invites explanation/justification of a contribution or speculation (new scenarios) /prediction/hypothesis. E.g. 'Why do you think the bottle floats?'	19.88 (10.89)
Reasoning (RE)	Provides an explanation or justification of own or another's contribution, or speculates, predicts, hypothesizes with grounds given. E.g. [After 'He came back'] 'because he made a promise'.	55.61 (16.21)
Invite co-ordination (CI)	Invites synthesis, summary, comparison, evaluation or resolution based on two or more contributions. E.g. 'Would anyone like to summarize the ideas we've been hearing?'	0.07 (0.37)
Simple co-ordination (SC)	Synthesizes or summarizes collective ideas (including own and/or others' ideas). Compares, resolves or evaluates different opinions, perspectives and beliefs. E.g. 'Some of you are talking about weight and some	0.49 (0.87)

	are talking about size; both matter – things float when they're light for their size'.	
Reasoned co-ordination (RC)	Compares, evaluates, resolves two or more contributions in a reasoned fashion. It includes all SC descriptors plus a counter-argument, reasoned rebuttal, two partial truths. E.g. 'We've been arguing about how much of personality is inherited; twin studies show conclusively it's 50%'.	0.12 (0.49)
Agreement (A)	Explicit acceptance of or agreement with a statement(s). E.g. 'Brilliant', 'Good', 'Yeah', 'Okay', 'I agree with X…'.	82.28 (25.37)
Querying (Q)	Doubting, full/partial disagreement, challenging or rejecting a statement. Includes a simple 'no' response when it shows rejection of an idea; not when in response to a question. E.g. 'Do you really think these angles are the same?'	20.28 (9.65)
Reference back (RB)	Introduces reference to previous knowledge, beliefs, experiences or contributions (includes procedural references) that are common to the current conversation participants. Includes *inviting* reference back. E.g. 'Can you share with us what we were just having a quick chat about please?'	6.09 (4.55)
Reference to wider context (RW)	Making links between what is being learned and a wider context by introducing knowledge, beliefs, experiences or contributions from outside of the subject being taught, classroom or school. Includes *inviting* reference to wider context. E.g. 'It's like in Macbeth where the storm builds into it'.	3.62 (3.31)
Other invitations (OI)	Invitations of all kinds of verbal contributions (e.g. opinions, ideas, beliefs), except for those coded as ELI, REI or CI. This includes invitations on a new topic if this does not fall in another invitation code, and procedural questions.	136.47 (34.59)

Un-coded (UC)	All turns not falling into at least one of the above categories. These typically include replies to OI questions, procedural remarks or spontaneously offered new ideas that do not appear to relate to any previous utterances or activities.	216.10 (75.91)

Table 2

Distributional Properties of Dialogue and Outcome Variables across Classrooms

Variable	Mean	SD	Range
Elaborated	205.72	78.04	85.99 to 430.30
Reasoned	153.09	50.91	64.33 to 280.31
Elaborated/Non-Dialogic Ratio	3.93	1.85	1.58 to 9.35
Reasoned/Non-Dialogic Ratio	5.32	2.68	1.37 to 14.01
Querying	40.95	19.95	9.51 to 86.90
Referring Back	6.09	4.55	0 to 26.24
Referring Widely	3.62	3.31	0 to 17.01
Aims	1.81	0.74	0 to 4
Monitoring	3.24	0.90	1 to 4
Reflection	1.36	1.26	0 to 4
Student Participation	2.46	1.07	0 to 4
Reading SAT	105.09	8.35	95.28 to 112.78
SPAG SAT	106.26	7.49	98.62 to 113.93
Mathematics SAT	104.59	7.09	97.92 to 113.28
Science	13.99	4.42	10.29 to 21.07
Reasoning	24.01	5.33	18.96 to 30.24
End-of-Year PASS	2.47	0.62	0.44 to 3.96

Table 3

Significant/Near-Significant Relations between Potential Confounds and Outcome Variables

Potential Confound	Outcome Variable	Significance
PreMaths (M)	Reading SAT	$t(72.40)=5.68$, $p<.001$
PreMaths (M)	SPAG SAT	$t(70.15)=2.78$, $p=.007$
PreMaths (SD)	Mathematics SAT	$t(60.07)=-2.41$, $p=.019$
PreRead (M)	Mathematics SAT	$t(58.12)=2.73$, $p=.008$
PreRead (M)	End-of-year PASS	$t(71.57)=1.98$, $p=.051$
PrePASS	Mathematics SAT	$t(62.30)=2.98$, $p=.004$
PrePASS	SPAG SAT	$t(71.29)=2.04$, $p=.045$
PrePASS	Science	$t(47.48)=1.98$, $p=.053$
Group	Mathematics SAT	$t(59.08)=2.60$, $p=.012$
Group	SPAG SAT	$t(68.50)=2.33$, $p=.023$
Group	Reasoning	$t(72.88)=2.53$, $p=.014$
Group	End-of-year PASS	$t(71.65)=-2.31$, $p=.024$

Table 4

Covariates (excluding Dialogue Variables) Used in Analyses of Dialogue-Outcome Relations

Outcome Variable	Baseline Index	Confound Variable
Mathematics SAT	PreMaths (M)	For Elaborated: PreRead (M); Group
		For Reasoned: PreRead (M)
		For Elaborated/Non-Dialogic: PrePASS
		For Reasoned/Non-Dialogic: PreRead (M); PrePASS
		For Querying: PreMaths (SD)
		For Referring Widely: PrePASS
		For Reflection: PreRead (M); PreMaths (SD)
		For Student Participation: Group
Reading SAT	PreRead (M)	For Elaborated: PreMaths (M)
		For Elaborated/Non-Dialogic: PreMaths (M)
		For Querying: PreMaths (M)
		For Monitoring: PreMaths (M)
SPAG SAT	PreRead (M)	For Elaborated: PreMaths (M); Group
		For Elaborated/Non-Dialogic: PreMaths(M); PrePASS
		For Reasoned/Non-Dialogic: PrePASS
		For Querying: PreMaths (M)
		For Referring Widely: PrePASS
		For Monitoring: PreMaths (M)
		For Student Participation: Group
Science	PreMaths (M)	For Elaborated/Non-Dialogic: PrePASS
	PreRead (M)	For Reasoned/Non-Dialogic: PrePASS

		For Referring Widely: PrePASS
Reasoning	PreMaths (M)	For Elaborated: Group
	PreRead (M)	For Student Participation: Group
End-of-Year PASS	PrePASS	For Elaborated: PreRead (M); Group
		For Reasoned: PreRead (M)
		For Reasoned/Non-Dialogic: PreRead (M)
		For Reflection: PreRead (M)
		For Student Participation: Group

Table 5

Significant/Near-Significant Relations between Dialogue and Outcome Variables

Dialogue Variable/s	Outcome Variable	Significance
Reasoned/Non-Dialogic Ratio x Refer Back	Mathematics SAT	$t(59.74)=-2.41$, $p=.019$
Elaborated x Student Participation	Mathematics SAT	$t(67.48)=3.40$, $p=.001$
Querying x Student Participation	Mathematics SAT	$t(66.08)=2.21$, $p=.03$
Elaborated x Student Participation	SPAG SAT	$t(76.13)=1.98$, $p=.052$
Querying x Student Participation	SPAG SAT	$t(74.95)=2.00$, $p=.049$
Elaborated x Refer Back	Reasoning	$t(79.00)=-2.05$, $p=.044$
Elaborated/Non-Dialogic Ratio x Refer Back	Reasoning	$t(71.67)=1.97$, $p=.052$
Elaborated	End-of-Year PASS	$t(71.54)=5.13$, $p<.001$
Elaborated/Non-Dialogic Ratio	End-of-Year PASS	$t(71.88)=-2.90$, $p=.005$
Reasoned/Non-Dialogic Ratio	End-of-Year PASS	$t(71.56)=-2.24$, $p=.028$
Elaborated/Non-Dialogic Ratio x Querying	End-of-Year PASS	$t(71.50)=2.30$, $p=.024$
Reasoned/Non-Dialogic Ratio x Querying	End-of-Year PASS	$t(71.43)=1.97$, $p=.052$
Reasoned x Querying	End-of-Year PASS	$t(71.51)=-1.79$, $p=.077$

Table 6

Estimated Marginal Mean SAT Scores for Classes Categorized via Elaborated, Querying and Student Participation

Class	Mathematics	SPAG	Reading
>Average EL+Q >Average SP	107.68_a	108.71_a	106.10_a
>Average EL+Q <Average SP	104.57_b	106.01_b	104.17_b
<Average EL+Q >Average SP	103.64_b	105.66_b	105.55_b
<Average EL+Q <Average SP	104.81_b	106.93_b	106.22_a
Significance	$F(3,1061)=11.88$ $p<.001$	$F(3,1164)=7.59$ $p<.001$	$F(3,1165)=3.48$ $p=.016$

Notes: a) Average = Mean score across classes; b) EL = Elaborated, Q = Querying, SP = Student Participation; c) When subscripts within columns differ, mean differences are statistically significant (Bonferroni, $p<.05$)

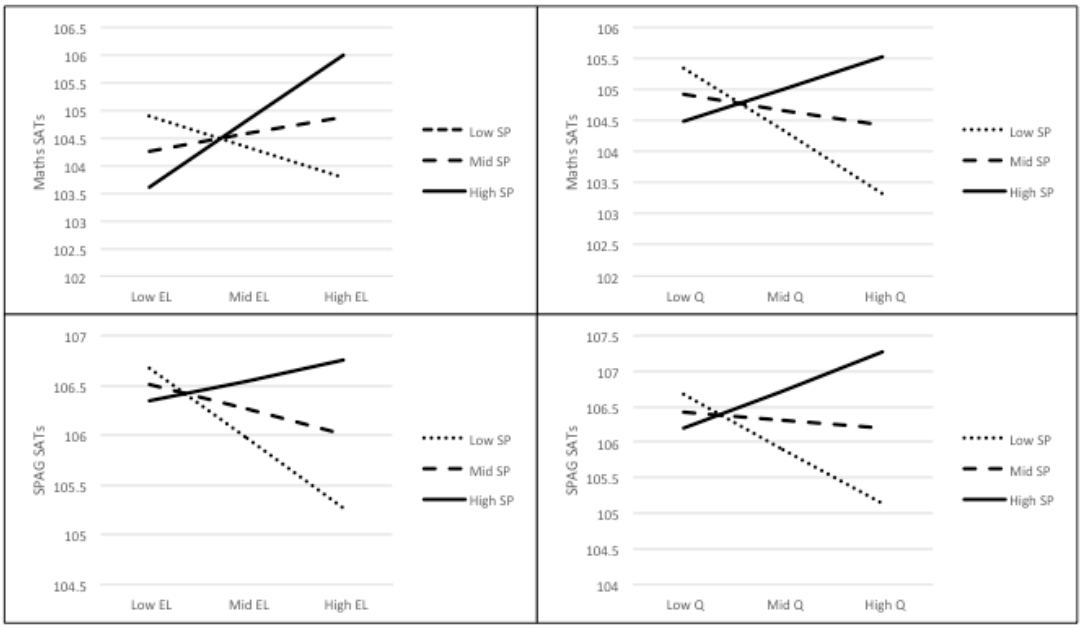

Figure 1. Interactions involving Student Participation as predictive of SAT mathematics and SPAG scores. SP= Student Participation. Q = Querying. EL = Elaborated.

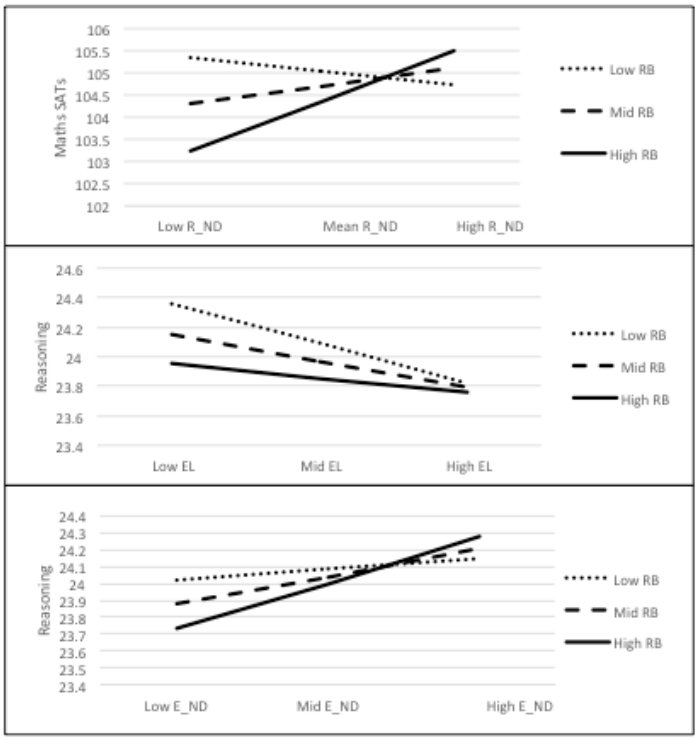

Figure 2. Interactions involving Referring Back as predictive of SAT mathematics and reasoning scores.
RB = Referring Back. R_ND = Reasoned/Non-Dialogic Ratio. EL = Elaborated. E_ND = Elaborated/Non-Dialogic Ratio.

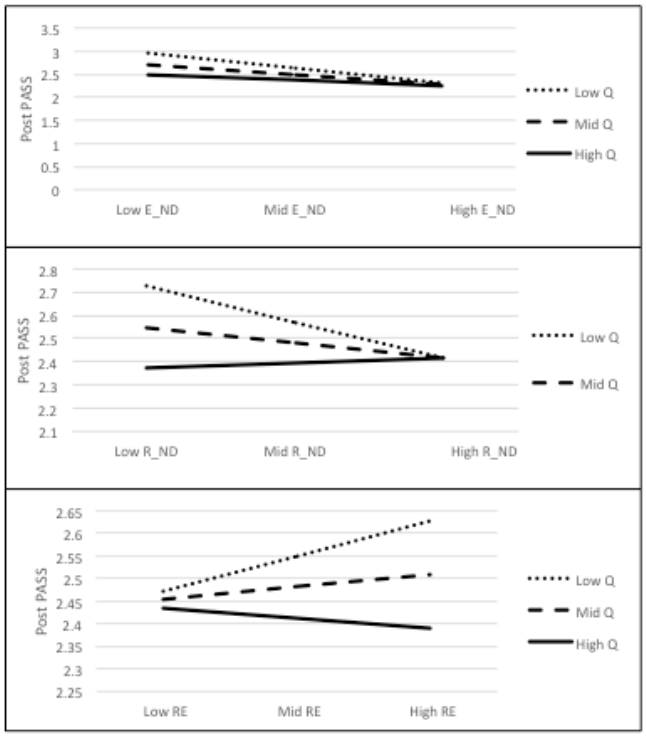

Figure 3. Interactions involving Querying as predictive of end-of-year pupil attitudes.
Q = Querying. E_ND = Elaborated/Non-Dialogic Ratio. R_ND = Reasoned/Non-Dialogic Ratio. RE = Reasoned.