EXPLORING THE INFLUENCE OF HOMOGENEOUS VERSUS HETEROGENEOUS GROUPING ON STUDENTS' TEXT-BASED DISCUSSIONS AND COMPREHENSION

P. Karen Murphy[1]

Jeffrey A. Greene[2]

Carla M. Firetto[1]

Mengyi Li[1]

Nikki G. Lobczowski[2]

Rebekah F. Duke[2]

Liwei Wei[1]

Rachel M. V. Croninger[1]

[1]*The Pennsylvania State University*

[2]*The University of North Carolina-Chapel Hill*

Published: October 2017

Author Note

This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A130031 to the Pennsylvania State University. Any opinions, findings, and conclusions or recommendations expressed are those of the author(s) and do not represent the views of the Institute or the U.S. Department of Education.

Correspondence regarding this article should be addressed to P. Karen Murphy, 102 CEDAR Building, Department of Educational Psychology, Counseling, and Special Education, The Pennsylvania State University, University Park, PA 16802

Contact: pkm15@psu.edu

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Abstract

Small-group, text-based discussions are a prominent and effective instructional practice, but the literature on the effects of different group composition methods (i.e., homogeneous vs. heterogeneous ability grouping) has been inconclusive with few direct comparisons of the two grouping methods. A yearlong classroom-based intervention was conducted to examine the ways in which group composition influenced students' discourse and comprehension. Fourth- and fifth-grade students (N = 62) were randomly assigned to either a homogeneous or heterogeneous ability small-group discussion. All students engaged in Quality Talk, a theoretically- and empirically-supported intervention using small-group discussion to promote high-level comprehension. Multilevel modeling revealed that, on average, students displayed positive, statistically and practically significant gains in both basic and high-level comprehension performance over the course of Quality Talk. Further, our findings indicated heterogeneous ability grouping was more beneficial than homogeneous ability grouping for high-level comprehension, on average, with low-ability students struggling more in homogeneous grouping. With respect to student discourse, additional quantitative and qualitative analyses revealed group composition differences in terms of the frequency, duration, and quality of student questions and responses, as well as the types of discourse low-ability students enacted in homogeneous groups. This study expands upon the extant literature and informs future research and practice on group composition methods.

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Exploring the Influence of Homogeneous Versus Heterogeneous Grouping on Students' Text-Based Discussions and Comprehension

Small-group activities and discussions are pervasive instructional practices in contemporary classrooms (Johnson, Johnson, & Stanne, 2000). Indeed, the prevailing instructional perspective seems to be that small-group activities and discussions promote enhanced learning, social engagement, and accountability (Slavin, 1991, 2011). For example, homogeneously grouping students by relative ability[1] or prior achievement allows teachers to adapt their instructional pace to accommodate the aptitudes or needs of particular groups (e.g., differentiated instruction; Coldiron, Braddock, & McPartland, 1987). This type of homogeneous ability grouping is particularly prominent in tiered literacy interventions (Torgesen et al., 2006). By comparison, arranging students into heterogeneous ability groups, as is common in text-based discussions, allows teachers to take advantage of student diversity and encourage collaboration among peers to enhance student learning and interdependence (Wilkinson, Soter, & Murphy, 2010).

The challenge, however, is that the functioning, productivity, and learning outcomes of small-group classroom discussions seem to vary by the group composition (e.g., homogeneous versus heterogeneous ability), goals (e.g., affective), and social and intellectual facilitation (e.g., teacher or peer) of the group (Azmitia, 1988; Lou et al., 1996; Saleh, Lazonder, & De Jong, 2005). Further, although predominant approaches to small-group, text-based discussions

[1] The term "ability" is often used in the grouping and discussion literatures, therefore we have used it in this article as well. However, "ability" in this sense does not mean a static or trait-like characteristic, rather it refers to measured ability at a particular point in time, which can and does change as a result of student and teacher effort.

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

exclusively encourage the use of heterogeneous ability groups, little is known regarding how group composition affects small-group discussions or learning from text (Murphy, Wilkinson, Soter, Hennessey, & Alexander, 2009). As such, the purpose of the present study was to examine the ways in which group composition influences students' text-based discussions and comprehension over time.

Ability Grouping Versus Whole-Class Instruction

Research findings have firmly established the benefits of small-group instruction as compared to whole-class instruction. In fact, a number of meta-analyses have been conducted to examine the effects of within-class grouping on achievement (Kulik, 1992; Lou et al., 1996; Slavin, 1987), all of which have overwhelmingly illustrated that grouped instruction was superior to whole-class or non-grouped instruction in promoting student learning. For example, Slavin (1987) reported a moderate advantage of within-class grouping over no grouping in upper elementary mathematics classes, especially when the number of groups was small (median *ES* = +.34). Similarly, Kulik (1992) reviewed eleven studies of within-class grouping from second to eighth grades and reported higher overall achievement levels in mathematics and reading for students grouped within classes, compared to their counterparts without grouping (mean *ES* = +.25).

A more comprehensive meta-analysis conducted by Lou et al. (1996) examined the results from 51 studies comparing the effects of grouping versus no grouping on achievement from first grade to college levels. The results revealed that within-class grouping positively influenced student learning in all content areas (mean *ES* = +.17) and that the grouping effect was statistically significantly greater in math and science (mean *ES* = +.20) than in reading, language arts, or other subject areas (mean *ES* = +.13). The results also showed that students of

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

varying ability levels (i.e., low, average, and high) all benefited from being assigned to small groups (mean *ES* = +.37, +.19, and +.28, respectively). Although low-, average-, and high-ability students differed in how much they benefitted from being assigned to small groups, the results showed that low-ability students gained statistically significantly more than average-ability students. Importantly, Lou et al. also explored the findings by examining the features of individual studies and found that differentiated instruction was more effective when provided in small groups (mean *ES* = +.25) than when the same instruction was provided as whole-class instruction (mean *ES* = +.02). Group size was also found to be linked to the grouping effect. Specifically, the effect size for small groups with three to four members (mean *ES* = +.22) was statistically significantly higher than for groups with five to seven members (mean *ES* = -.02).

Homogeneous Versus Heterogeneous Grouping

While the superiority of within-class ability grouping is undergirded by a wealth of research, there appears to be no single best evidence-based practice for creating small groups, particularly when the goal is to enhance text-based discussion and comprehension. The notable exception is that individual differences in students' domain-general ability (e.g., intelligence) or domain-specific ability (e.g., reading competence) are almost always taken into consideration in group creation within classrooms. Indeed, the most controversial issue underlying group composition is whether small groups should be comprised of students who are of similar (i.e., homogeneous) or dissimilar (i.e., heterogeneous) ability levels. In the meta-analysis by Lou et al. (1996), 12 of the reviewed studies compared the effects of homogeneous grouping to heterogeneous grouping on student achievement and suggested a result favoring homogeneous grouping (*ES* = +.12, $p < .05$). However, the advantage of homogeneous grouping was not uniform across students of different ability levels. Specifically, low-ability students were found

to learn more in heterogeneous groups (*ES* = -.60, $p < .05$), average-ability students gained more in homogeneous groups (*ES* = +.51, $p < .05$), and high-ability students performed equally well in either group, regardless of ability composition (*ES* = +.09, stat ns). Lou et al. also found that subject area was a statistically significant moderator of the effects of group composition on student achievement. Among the findings summarized in the meta-analysis, only four compared the effects of group composition in reading and these findings revealed a medium effect size favoring homogeneous grouping (*ES* = +.36, $p < .05$). By contrast, the effect of group composition was not statistically significantly different from zero in math and science.

The findings reported in Lou et al. (1996) are also supported by a number of individual studies not included in the research synthesis (e.g., Azmitia, 1988; Saleh et al., 2005; Webb, 1980, 1991). Among them, Saleh et al. (2005) examined how group composition influenced students' achievement, social interaction, and motivation in a biology course. A total of 104 fourth-grade students were identified as being of relatively low, average, or high ability based on their scores on a standardized science test and then randomly assigned to one of 13 homogeneous groups (i.e., four low-, five average-, and four high-ability groups) or 13 heterogeneous groups, each with four students (i.e., one low-, two average-, and one high-ability student). All groups received the same instruction over the course of 16 plant biology lessons, which included brief whole-class instruction at the beginning followed by collaborative learning tasks. The results showed that low-ability students in heterogeneous groups performed better on the individual posttest and were more motivated to learn compared to their low-ability peers in homogeneous groups. Average-ability students seemed to benefit more from learning in homogeneous groups, as compared to heterogeneous groups, and high-ability students exhibited equally strong learning outcomes regardless of their membership in either homogeneous or heterogeneous groups.

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Importantly, Saleh et al. (2005) also examined the social interaction in both grouping conditions and discovered that heterogeneous grouping elicited higher proportions of individual elaborations (i.e., elaborations made by a single student), whereas homogeneous grouping triggered more co-construction of elaborations (i.e., elaborations constructed across multiple students). Indeed, group composition not only affects students' academic attainment but also exerts influence on students' social interactions. These social interactions may be an important mediator of the effect of group composition on small-group learning (Saleh et al., 2005; Webb & Palincsar, 1996). This finding aligns with both Piaget's and Vygotsky's theories on learning and development. According to Piaget (1932), interacting with peers forces students to recognize the gaps or contradictions in their understanding, helps them to repair misconceptions, and develops their more advanced cognitive architecture. Thus, working with more competent peers is likely to stimulate more cognitive conflict than working with similar-ability peers. According to Vygotsky (1978), social interaction is optimal for children's cognitive development when collaborating with someone of higher ability. With the assistance provided by a more capable peer, children gradually internalize the skills above their current developmental level so that they can perform the tasks independently. Hence, small groups provide students with opportunities to engage in social interaction with peers, which has an important influence on their achievement and social participation (Rosenbaum, 1980; Wilkinson & Fung, 2002).

Additionally, these theoretical notions provide insights into the differential effects of group composition on student learning. In particular, these theoretical premises suggest why low-ability students benefit more by learning in heterogeneous groups with higher-ability peers than in homogeneous groups with only low-ability peers. Indeed, research on group processes has found that low-ability students tend to exhibit more help-seeking behaviors and thus receive

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

more explanations and support in heterogeneous groups than in homogeneous groups (Azmitia, 1988; Tudge, 1989; Webb, 1980). Hearing the elaborated explanations of peers enables low-ability students to fill in knowledge gaps and correct their own misconceptions. Low-ability students who work with higher-ability peers are also more likely to be exposed to reasoning skills that they do not currently possess (Tudge, 1989). In contrast, decades of research on group learning has failed to show achievement benefits for students in homogeneous low-ability groups (Allington, 1980; Barr, 1975; Lou et al., 1996; Rosenbaum, 1980; Slavin, 1987). Studies have shown that placing students into low-ability groups greatly increased the likelihood that they would become inattentive during group work and substantially increased the role of teachers in enforcing behavioral norms (Eder & Felmlee, 1984). The aforementioned findings are compounded by the fact that teachers establish differential expectations when students are homogeneously grouped by ability (Metz, 1978). In particular, inattentive behaviors were more tolerated in homogeneous low-ability groups compared to other groups, and perhaps relatedly, homogeneous grouping was also shown to have negative effects on low-ability students' self-concepts (Rosenbaum, 1980).

While consistent findings have shown that heterogeneous grouping is more beneficial for low-ability students, high-ability students' performance is generally unaffected by group composition (Lou et al., 1996; Saleh et al., 2005). However, there is evidence to suggest that the role and performance of high-ability students in small groups is somewhat affected by the constitution of the group. For example, Webb (1980, 1991) found that high-ability students in heterogeneous groups tended to adopt the role of teacher or leader and provide more elaborated explanations to other group members, especially the low-ability students (Webb, 1980, 1991). Johnson, Skon, and Johnson (1980) also found that high-ability students developed more

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

sophisticated reasoning strategies when working in heterogeneous groups than in homogeneous groups as they had more opportunities to teach others. Conversely, when high-ability students participated in homogeneous groups, they were more likely to co-construct knowledge by elaborating on one another's ideas and producing more collaborative reasoning (Fuchs, Fuchs, Hamlett, & Karns, 1998; Webb, Nemer, Chizhik, & Sugrue, 1998).

As for average-ability students, many researchers expressed concerns that these students do not take full advantage of learning in heterogeneous groups (e.g., Saleh et al., 2007; Webb & Palincsar, 1996). Arguably in heterogeneous ability groups, high-ability students and low-ability students tend to form a teacher-learner relationship, leaving fewer opportunities for average-ability students to offer or receive help and explanations than what might be available in homogeneous groups (Webb & Palincsar, 1996). Saleh et al. (2007) suggested that the establishment of ground rules and structuring of group roles might be effective in supporting average-ability students' contribution to discussion and promoting their achievement, motivation, and engagement.

Taken together, it seems that no singular form of group composition is equally advantageous for all students. Meta-analytic data showed a slight advantage of homogeneous over heterogeneous grouping in reading, but this finding was based on a small number of studies (Lou et al., 1996). The scant research evidence is mixed, but there is some indication that heterogeneous grouping may be more beneficial for low-ability students (Rosenbaum, 1980), whereas homogeneous grouping may be more beneficial for average-ability students (Fuchs et al., 1998; Webb et al., 1998). These studies have examined effects of the type of group composition on achievement, but there is some evidence that these effects are mediated by the nature of the social interactions within these groups (Saleh et al., 2005). Clearly, more research

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

is needed to examine the effects of within-class ability grouping on both the nature of text-based discussions as well as students' concomitant reading achievement (Wilkinson & Fung, 2002).

Small-Group Discussions to Enhance High-Level Comprehension of Text

In the present study, we investigated the effect of homogeneous and heterogeneous grouping on students' small-group discussion as well as basic and high-level comprehension of text. However, productive discourse does not naturally result from simply putting students in groups. Research has shown that small-group classroom discourse can promote basic comprehension, but only particular kinds of discourse are likely to promote the kind of high-level comprehension and higher-order critical-analytic thinking necessary for effective democratic participation (Murphy et al., 2009). Indeed, *high-level comprehension* is a requisite skill for students to learn (Common Core State Standards, 2011); it stipulates that students can critically and reflectively consider different perspectives about, around, and with text (i.e., critical-analytic thinking) and meaningfully evaluate the nature and quality of the arguments or information in the text (i.e., epistemic cognition; Greene, Sandoval, & Bråten, 2016; Murphy, 2007). Unfortunately, a majority of American students struggle with acquiring basic comprehension of text, let alone high-level comprehension (National Assessment Governing Board [NAGB], 2015).

Factors influencing high-level comprehension. In an effort to address the aforementioned problem, researchers have identified several predictive factors, namely basic comprehension, prior knowledge, oral reading fluency, and epistemic cognition, that are essential to high-level comprehension. These predictors provide important guidelines in terms of developing instructional approaches aimed to promote students' high-level comprehension. First, students need basic comprehension about the text as a foundation to think critically and

analytically around and with the text while developing high-level comprehension (Bråten, Britt, Strømsø, & Rouet, 2011; Kintsch, 1998). Second, students' high prior knowledge can facilitate and support both basic and high-level comprehension by enhancing text recall (Alexander, Kulikowich, & Schulze, 1994), generation of inferences (Kendeou & van den Broek, 2007; McNamara, Kintsch, Songer, & Kintsch, 1996), and engagement with the text (e.g., providing more questions and elaborated responses in text-based discussions; Goatley, Brock, & Raphael, 1995). Third, efficient word recognition as indicated by oral reading fluency is crucial in terms of providing sufficient capacity for high-level comprehension (Fuchs, Fuchs, Hosp, & Jenkins, 2001). Finally, students must develop effective epistemic cognition to obtain, comprehend, and apply knowledge (Greene et al., 2016) and actively engage in epistemic practices to achieve high-level comprehension. In particular, students need to perceive texts as constructed and reflectively evaluate arguments in the text as opposed to accepting the textual information as given facts (Bråten et al., 2011).

Small-group discussions promoting high-level comprehension. Recently, concerted efforts have been devoted to investigating small-group, text-based discussion as an effective approach to promoting students' reading comprehension (Murphy et al., 2009). However, it should be noted that whereas classroom discussions have been found to augment students' basic comprehension (McKeown, Beck, & Blake, 2009), only certain types of talk or discourse propel discussions into an epistemic mode: a mode requisite for promoting high-level comprehension (Murphy et al., 2009). An exhaustive meta-analysis conducted by Murphy et al. (2009) identified nine discussion approaches aimed at promoting high-level comprehension and evaluated the differential contributions of each identified discussion approach. Among the various stances (i.e., efferent, expressive, and critical-analytic) that these discussion approaches

held towards text, efferent approaches better supported students' basic comprehension by focusing on the retrieval of information in the text. Alternatively, critical-analytic approaches contributed more to high-level comprehension by allowing students to query the text and to evaluate multiple perspectives. In essence, critical-analytic approaches were more likely to cultivate effective epistemic cognition and appeared to be more effective at promoting high-level comprehension, compared to efferent or expressive approaches. Further, a consistent message that can be gleaned from the reviewed studies is that specific instructional practices promote learning within and from these groups (Kulik & Kulik, 1987, 1992; Lou, Abrami, & Spence, 2000). Simply put, instructional practices in which students receive explicit tasks or lessons on productive discourse have been shown to bolster the achievement effects of within-class grouping (Lou et al., 1996). In addition to being explicit, instructional practices must also leverage the aforementioned predictive factors to optimize small-group discourse effects on students' high-level comprehension of text.

Wilkinson et al. (2010) used the findings from Murphy et al. (2009) to develop Quality Talk, a small-group discussion approach that integrated the best features of various discussion approaches to foster high-level comprehension of text. Theoretical foundations that support the use of Quality Talk in stimulating reading comprehension have many facets, including cognitive, sociocognitive, social constructivist, and dialogic perspectives. The cognitive nature of classroom discussion stipulates that students must cognitively engage in the construction of meaning during discussion (McKeown et al., 2009), while the sociocognitive foundation supports the use of discussion so that students can express their own voices while hearing others' opinions and ideas. From the perspective of social constructivists, small-group discussions allow students to use talk as a tool to co-construct knowledge and advance thinking, while also

promoting the sharing of control between a more knowledgeable other (e.g., teacher or peers) and students in the group (Vygotsky, 1978). Finally, supported by the dialogic point of view, the conflicting voices raised during a small-group discussion can augment students' comprehension (Wilkinson et al., 2010).

Essentially, Quality Talk gives prominence to fostering students' epistemic cognition by encouraging students to ask deep, meaningful questions about the text and to reflectively and critically consider the quality of the arguments and make reasoned judgments (i.e., critical-analytic stance). In addition, students are also encouraged to connect their personal experience with the text (i.e., expressive stance) and to retrieve information from the text (i.e., efferent stance). As such, students can more effectively construct basic comprehension of the text and activate their prior knowledge by making connections to the text being discussed. Notably, the shared control between teachers and students in Quality Talk supports the emphasis on both expressive and efferent stances and also ensures that teachers can choose texts that are rich and interesting enough to discuss. In Quality Talk, teachers are expected to release increasing responsibility and interpretative authority to students across a series of discussions, as students take on more responsibility within their discussion groups and interact more closely and frequently with their peers. This is achieved through a series of explicit, discourse mini-lessons on questioning and argumentation combined with practice and implementation of key skills in small-group discussions (see www.qualitytalk.psu.edu for sample mini-lessons). What is inconclusive in the literature is how group composition can moderate the effect of small-group discussion on students' discourse and subsequent basic and high-level comprehension of text.

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Research Questions

The current investigation is an extension of the research on the effects of within-class ability[2] grouping on student discourse and reading achievement. Specifically, we explored the influence of homogeneous and heterogeneous ability group composition on classroom discourse and individual reading outcomes, including high-level comprehension, with the implementation of the Quality Talk (QT) intervention under ecologically valid conditions over the course of a yearlong intervention. A number of specific research questions (RQ) guided this investigation including:

RQ1. To what extent do students' basic and high-level comprehension of text change over the course of the Quality Talk intervention, controlling for text and topic knowledge, and are these changes moderated by grade or oral reading fluency scores?

RQ2. To what extent do changes in comprehension of text, over the course of the Quality Talk intervention, as evidenced in students' performance on the basic and high-level comprehension measures, vary by the nature of the group composition (i.e., homogeneous vs. heterogeneous)?

RQ3. To what extent does homogeneous and heterogeneous ability group discussions vary with respect to *students'* discourse elements and *teachers'* use of discourse moves?

RQ4. In what ways do the experiences of low-ability students differ across types of grouping and differ from their high-ability peers?

---

[2] As one reviewer noted in a previous version of the manuscript, *ability* might more precisely be operationalized as *initial performance level*, yet in order to maintain consistency with the extant literature we continue to employ the term ability throughout.

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Method

Participants

Four teachers and their 4th- and 5th-grade students from one elementary school were recruited to participate at the beginning of the school year. With the exception of two students, all parents consented and all students assented to participate in the research. One 4th-grade student left the school shortly after the beginning of the school year. Thus, participants included 62 students from both 4th grade (n = 28; female = 15) and 5th grade (n = 34; female = 19). The school served approximately 300 students (30% free or reduced lunch) from kindergarten through fifth grade, and it was located in a small, Midwestern city. The students at the school were predominantly Caucasian (86%); however, a few students identified themselves as American Indian/Alaska Native (2%), Asian (2%), Black (2%), Hispanic (2%), and a small percentage identified with more than one racial group (5%).

Grouping

As will be described later, students' oral reading fluency scores were collected at the beginning of the year and were used to determine grouping assignment (i.e., homogeneous groups of students with similar ability levels *or* heterogeneous groups of students with wider variations in ability levels). We chose oral reading fluency as the grouping variable instead of other potential variables such as prior knowledge or epistemic cognition because it is a standardized, curriculum-based measure commonly employed in elementary schools. Moreover, multiple studies have identified oral reading fluency scores as the most valid indicator of student reading comprehension ability (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Goffreda & DiPerna, 2010; Johnson, Jenkins, Petscher, & Catts, 2009). For example, among the measures assessed in DIBELS (i.e., Dynamic Indicators of Basic Early Literacy Skills; a widely-used screening and

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

placement tool in the primary grades), Johnson and colleagues (2009) singled oral reading fluency out as having the highest classification accuracy.

The teachers agreed to group the students across the entire grade in order to achieve optimal grouping as well as to alleviate the nestedness of students within teacher. A similar procedure was employed across both grades. Within each grade, students were ranked based on their oral reading fluency (i.e., number of words read correctly) and paired with a similar-ability, grade-level peer. Each student was assigned a number using the random number generator in Excel, and the student with the lower number of each pair was assigned to a heterogeneous group and the other student assigned to a homogeneous group. Thus, in each grade the third of students with the highest oral reading fluency across both classes were split between the homogeneous high-ability group and the three heterogeneous groups, the middle third of the students were split between the homogeneous average-ability group and the three heterogeneous groups, and the lowest third of the students were split between the homogeneous low-ability group and the three heterogeneous groups. In this way, the composition of the groups maintained the highest degree of homogeneity within homogeneous groups while also maximizing the heterogeneity across each of the three heterogeneous groups with respect to students' oral reading fluency abilities. Further, by grouping across class and within grade, we removed a potential teacher confound; all groups contained students from both classes and over time teachers facilitated all groups.

Initially in fourth grade, the students assigned to homogeneous groups were equally split into three groups based on their ranked order of oral reading fluency: low-ability, average-ability, and high-ability groups. Because the number of students in 5th grade could not be equally split into three groups, the homogeneous group composed of all high-ability students was assigned one fewer student than the other two homogeneous groups. The students assigned to

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

heterogeneous grouping were placed in one of the three groups by alternating assignment based on their ranked order. Thus, the lowest three students assigned to heterogeneous groups were distributed across each of the three heterogeneous groups, continuing distributing students with increasing ability levels across the three groups, such that each heterogeneous group contained students with low-, average-, and high-fluency abilities.

Once the randomized grouping assignments were made, group oral reading fluency means and standard deviations were calculated (e.g., to ensure homogeneous groups had lower standard deviations than heterogeneous groups) and the groups were hand checked to ensure that they contained a mix of students from each class, along with a mix of genders. Grouping was slightly modified based on these calculations and checks. For example, when a group contained students of all one gender or all one class, grouping was modified by reassigning similar ability students from one group into another. Once the researchers finished configuring the groups, teachers were consulted to finalize the grouping composition. For both grades, students were grouped into three homogeneous groups and three heterogeneous groups (i.e., six groups in 4th grade and six groups in 5th grade). In fourth grade, one student left the school shortly after baseline. Because that student was assigned to the homogeneous high-ability group, after the second discussion that group consisted of four students. Thus, the fourth-grade groups each contained four or five students; the 5th grade groups each contained either five or six students.

Grouping students across class also mitigated the likelihood of the teachers influencing the study outcomes. Within grade, the discussion groups were assigned such that each teacher facilitated three groups, including at least one homogeneous and one heterogeneous group. Teachers facilitated the same three groups' discussions (i.e., one homogeneous and two heterogeneous or vice versa) three to four times. After that, teachers continued to switch which

groups they facilitated every three to four times. By doing this, we decreased the likelihood that group composition and teacher effects would be confounded.

Intervention

The Quality Talk model is comprised of four components: an ideal instructional frame, discourse elements that promote high-level comprehension, pedagogical principles, and a set of teacher discourse moves. Teachers learned about the components of QT through initial and ongoing professional development. Although students did not learn about the teacher-specific components of QT, they were given explicit mini-lessons about how to productively participate in discussions by asking meaningful, authentic questions and how to come to an examined understanding by creating and weighing reasoned arguments. Student comprehension processes were further enhanced through a literacy journal in which students completed prediscussion and postdiscussion activities. Importantly, all QT discussions were guided by normative expectations that were communicated to students through a set of discussion ground rules, and all discussions ended with a debrief in which members established content- and process-related goals for the next discussion.

Professional development. An initial, two-day professional development workshop was provided to participating teachers before the implementation of Quality Talk. During the professional development, the teachers were introduced to the QT model. They were also taught how to conduct QT discussions effectively, deliver QT mini-lessons (i.e., questioning and argumentation lessons), and use effective discourse moves. In addition to the initial professional development, five discourse coaching sessions were conducted to provide continuous support and training. For each coaching session, teachers reviewed a video recording of one of their previously conducted discussions and completed the Discourse Reflection Inventory for

Teachers (DRIFT), a semi-structured tool designed to assist teachers to code, reflect on, and understand their discussions. As teachers completed the DRIFT, they recorded the turn-taking pattern of the discussion, identified the discourse elements present in their students' talk, and assessed their progress toward pre-established goals. After they completed the DRIFT, teachers met individually with a discourse coach, reviewed their discussions and the DRIFT, and established new goals and methods for continued success.

Quality Talk mini-lessons. Ten Quality Talk mini-lessons (i.e., six questioning lessons and four argumentation lessons) were developed by the researchers and provided to the four teachers during professional development. The Quality Talk questioning mini-lessons aimed to teach students how to produce different types of authentic questions (e.g., uptake questions, speculation questions, analysis questions, generalization questions, or connection questions) that elicit multiple possible answers while promoting rich discussions. The Quality Talk argumentation mini-lessons aimed to teach students how to respond to authentic questions using effective argumentation skills. Students were introduced to different components of argumentation (i.e., claim, reason, evidence, counterargument, and rebuttal). Students were encouraged to support their responses with reasons and evidence and to prompt others to do so by asking, "Why do you think that?" and "How do you know that?" Practice materials and animated videos were created along with the slides and lesson plans such that the teachers could use them to deliver Quality Talk mini-lessons with high implementation fidelity.

Quality Talk literacy journal. A researcher-designed Quality Talk literacy journal was developed for participants to use throughout the QT intervention. The journals contained pages for students to consider the text genre, main idea and supporting details, and to generate various kinds of authentic questions before participating in the QT discussion. Students were also

provided with space to respond to a story-related, authentic question in writing after the QT discussion. Journals were reviewed by teachers before and after discussions.

Quality Talk discussions. Weekly small-group discussions were conducted by teachers on the main reading selection from the school language arts curriculum (i.e., *Scott Foresman Reading Street*©). At 20 time points (i.e., Baseline and 19 QT discussions) over the entirety of the QT intervention, each discussion was facilitated by one of the two grade-level teachers and lasted about 15 to 20 minutes. Teachers conducted Baseline discussions with students in their classroom in a business-as-usual manner (e.g., whole class or small group). Each QT discussion group consisted of four to six students in total, either composed homogeneously or heterogeneously as determined by students' random assignment. See Appendix A for an excerpt from a fifth-grade Quality Talk discussion at midyear. Every QT discussion was video recorded and audio recorded.

Measures

Oral reading fluency. Students' oral reading fluency was assessed using the AIMSweb Reading Curriculum-Based Measure and was adopted as the criterion to determine grouping before the intervention. AIMSweb was administered at Baseline, Time 2, and Time 3; for this study, only Baseline data will be used. The standardized assessment evaluates the number of words read correctly in one minute (Shinn & Shinn, 2002). Specifically, students were assessed individually by a researcher who was trained in the standardized administration procedure. Students were asked to read aloud each of three passages for one minute. The researcher recorded a score for each passage by subtracting the number of skipped or mispronounced words from the total number of words read by the student for each corresponding passage. The researchers then gave the student a final score for oral reading fluency using the median score

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

across the three passages. Ample evidence for the reliability and validity of scores from the AIMSweb measure has been established in the literature (Shinn, Good, Knutson, Tilly, & Collins, 1992).

Discourse coding. Video recorded QT discussions for each group were coded by two trained discourse coders in accordance with a discourse coding manual adapted from Soter, Wilkinson, Murphy, Rudge, and Reninger (2006); see Murphy, Firetto, Greene, and Butler (2017) for the most recent version of this coding manual. Considering the varied durations of the discussions and content irrelevant to the discussion (e.g., review of ground rules or discussion debrief), the middle 10-minute segment was coded for the discussions conducted at Time 1 (i.e., intervention week 2), Time 2 (i.e., intervention week 10), and Time 3 (i.e., intervention week 19). Following the rules and procedures described in Murphy et al. (2017), coders reviewed and coded each 10-minute segment using *StudioCode*© (version 5.8.3) by: (a) identifying question events (i.e., an initiating question and all subsequent responses to that question) while also noting whether the initiating questions were asked by the teacher or a student, (b) coding the question type (i.e., authentic question or test question) of each question event based on the discourse within the event, (c) coding students' argumentation responses within question events (i.e., elaborated explanations and exploratory talk), and (d) identifying instances of teachers' use of discourse moves. See Appendix A for a transcribed excerpt of coded discourse. Inter-rater agreement was established by comparing an individual's codes with the reconciled codes (i.e., codes agreed upon by the two coders after resolving discrepancies and disagreements) for each discourse coder (see also Li, Murphy, Wang, Mason, Firetto, Wei, & Chung, 2016). After both coders achieved an inter-rater agreement above 80%, videos were coded independently. Coding

was periodically checked to ensure coders maintained agreement over 80% across the coded videos.

Reading comprehension measure. In order to assess students' basic and high-level comprehension using a standardized assessment, a reading comprehension measure was developed and administered at three time points (i.e., Baseline, Time 2, and Time 3). Three forms of parallel assessments were created based on three comparable reading selections (i.e., Text A, Text B, and Text C) identified from *We Were There, Too! Young People in U.S. History* by Phillip Hoose. The three selections shared similar grade-level difficulty (i.e., Flesch-Kincaid levels between 6.1 and 7.3), genre (i.e., biographical), and richness (i.e., representing different views and allowing for argumentation). The texts were modified for consistent sentence, paragraph, and text length (i.e., approximately 1000 words).

The corresponding assessments for each text were developed by the researchers in accordance with a table of specifications derived from the NAEP (2015) framework and in following the guidelines provided by Popham (2006). Each member of the assessment development team, comprised of a subset of the full research team, wrote items based upon each text, which were then pilot tested by other members of the team to evaluate each item's clarity and difficulty. Candidate items were refined and then the assessment development team created a draft version of the assessment from the item pool. Separate members of the research team then pilot tested the assessment, and the assessment was also vetted by classroom teachers. Feedback from these groups was integrated into another round of revision by the assessment team, producing final versions of each assessment.

Each form of assessment contained five multiple-choice questions that aimed to measure students' basic comprehension (i.e., locate/recall and integrate/interpret) and one written

argumentation assessment that aimed to measure students' high-level comprehension (i.e., critique/evaluate). The delivery of the three parallel assessments was counterbalanced to account for text effect. Specifically, each student was assigned a particular order of the text (e.g., Baseline, Time 2, Time 3 = A-B-C) for taking the comprehension measures. After approximately twenty minutes reading the assigned text, students were given fifteen minutes to complete the written argumentation assessment and then ten minutes to complete the five-item, multiple-choice assessment.

Each text told a story related to a different topic in United States history, including the Cherokee nation's adoption of written language, the challenges of the Great Migration, and the role of women in the Civil War. The written argumentation prompts included: "Which had a greater impact on the Cherokee culture: being moved from their lands or creating a written language?" "If you were Charles, where would you have wanted to live: Alabama or Detroit?" and "Should the doctor have kept Deborah's secret after she recovered so that she could have stayed in the army?" with each question followed by the instruction to "Provide reasons and evidence to support your answer." Written argumentation responses were scored according to a rubric, where the researchers assigned predetermined point values for the various components of argumentation (i.e., claim=1pt, reason=1pt, evidence=1pt, counter-argument reason=1pt, and rebuttal=1pt). Two doctoral student research assistants working on the project independently rated one-third of the written argumentation responses based on the rubric with 89% agreement. They discussed all the discrepancies until agreement was reached, and then the remaining essays were scored independently.

The necessity of creating new assessments tailored to these texts, and amenable to our teachers, meant that there was no pre-existing reliability or validity evidence available. Our

sample was too small to conduct a factor analysis for construct validity evidence and, indeed, was small enough that reliability estimates would also be questionable (Crocker & Algina, 1986). Further, our counterbalancing, to account for text effects, introduced time as another confound in our data. Therefore, we were unable to conduct extensive psychometric analyses. Nonetheless, we felt confident in our assessment development process, following the NAEP framework and Popham's (2006) guidelines.

Topic knowledge measure. Prior to reading the text for the reading comprehension measure, students received a text-specific topic knowledge measure. Thus, three forms of topic knowledge measures were developed for the three corresponding texts. The topic knowledge measure contained a 6-point response scale to measure how much the students thought they knew about the topic of the text (i.e., 1=Relatively Nothing, 6=A Great Deal) and an open-ended prompt that asked for students to recall words, phrases, or sentences related to the topic of the text. Students were given five minutes to complete the topic knowledge measure. Student responses to the prompt were segmented into idea units and each idea unit was given one point for relevance to the topic or no points if it was irrelevant or inaccurate. Two trained doctoral student research assistants developed banks of relevant idea units prior to reading student responses, as guides for topic knowledge scoring. Then the two raters compared their ratings on one-third of the responses, with interrater agreement of 89%, and reconciled disagreements through discussion. Then the raters scored the remaining responses independently. The 6-point response scale proved uninformative in initial analyses, therefore we utilized the idea unit topic knowledge measure only for all analyses. There were no statistically significant differences in topic knowledge scores by text at any of the three time points (all *p*s > .05).

Teacher feedback measure. A teacher feedback measure was administered following Time 2 and Time 3 to collect information pertaining to the usability, feasibility, and fidelity of the Quality Talk intervention. The feedback measure consisted of items eliciting teachers' views on QT in general, time it took to implement QT, QT mini-lessons, grouping, professional development, assessments, and their knowledge of QT. Teachers responded to the items with a Likert-type 6-point scale where 1 indicated Strongly Disagree and 6 indicated Strongly Agree. Additionally, the survey contained open-ended questions about QT, including coaching sessions, professional development, and the types of group composition.

Procedure and Analysis

Teachers, parents, and students were consented or assented, as appropriate, at the beginning of the school year. As part of the study, teachers conducted 20 discussions (i.e., Baseline and 19 Quality Talk discussions) over the course of the school year, and the study was comprised of four data collection time points: Baseline-preintervention, Time 1-intervention week 2, Time 2-intervention week 10, and Time 3-intervention week 19. Our quantitative analyses utilized all individual student outcomes and included measures of students' oral reading fluency, basic comprehension, and high-level comprehension at Baseline, Time 2, and Time 3. The classroom discussions conducted at about intervention weeks 2, 10, and 19 were analyzed both quantitatively and qualitatively. Baseline discussions were not examined because the students were grouped for the first Quality Talk discussion based on their oral reading fluency outcomes collected *at* Baseline, therefore there was no Quality Talk discussion data at Baseline and only Time 1, Time 2, and Time 3 were included in the discourse analyses.

Teachers began implementing the Quality Talk model as part of their language arts curriculum immediately after Baseline data collection and continued throughout the remainder of

the school year. As part of the intervention, teachers delivered the content of all questioning and argumentation mini-lessons. By grade, teachers conducted discussions on the same texts, based on their extant language arts reading curriculum. Mini-lesson instruction and QT discussions were video recorded and audio recordings were collected as a backup for fidelity and data purposes.

Data sources. To address RQ1 and RQ2, we analyzed individual student outcome data from 4th and 5th grade, collected at Baseline, Time 2, and Time 3, using piecewise multilevel modeling (Singer & Willet, 2003). Our analyses revealed the need for a more fine-grained exploration of the discourse. Thus, RQ3 and RQ4 were addressed using a subset of data from the fourth-grade discussions at Time 1, Time 2, and Time 3. Based on the nature of these research questions, we conducted a quantitative analysis of student-initiated discourse elements and teacher discourse moves for RQ3, examining three discussion groups (i.e., homogeneous low-ability group, homogeneous high-ability group, and one heterogeneous group). Discussions for these three groups were facilitated by the same teacher at Time 1, Time 2, and Time 3. We felt that this group selection would allow us to address the research question while helping to control for teacher variations. For RQ4, we initially chose to analyze discourse data from these same three groups at these same three time points. After initial qualitative coding of these three groups, however, we felt that the data were not saturated. One of the high-ability students in our heterogeneous group was very reserved and did not participate much in the discussion, whereas the other high-ability student in the group demonstrated behaviors similar to students in the homogeneous high-ability group. We determined that it would be worthwhile to explore additional data from another heterogeneous group to observe how high-ability students behaved in that group. Due to an equipment malfunction at intervention week 2, we only had all three

discussions video recorded for one of the remaining two heterogeneous groups. The heterogeneous group with all three video recordings was chosen to be added to the RQ4 qualitative analysis. In total, 12 transcriptions and videos (i.e., four groups across three time points) were analyzed for RQ4.

Usability, feasibility, and fidelity data. Given that our research questions were explored within the context of a small-group discussion intervention, Quality Talk, it was important to assess the degree to which the intervention was implemented with fidelity (Greene, 2015). Researchers reviewed video recordings of the Quality Talk mini-lessons from each of the four teachers in the study. They found high adherence (i.e., 96%) to the lesson plans provided to the teachers. These reviews, along with the on-going interactions with the teachers through coaching, supported a high level of implementation fidelity of Quality Talk. This was also supported by results from the teacher feedback measure. The researchers asked teachers about the effectiveness of each lesson as well as their knowledge of each lesson. The teachers' responses clearly reflected successful implementation and understanding of the different components of Quality Talk. Specifically, each lesson received an average of at least 4.75 out of 6 for effectiveness and 5.5 out of 6 for teacher understanding. This feedback also served as a measure of the interventions' overall usability and feasibility. Overall, all four teachers indicated a positive experience in regards to Quality Talk. They generally felt that it was an effective instructional strategy for their students and a good fit within their school environment. All teachers strongly agreed that Quality Talk was effective in fostering critical thinking in language arts, while three of the four teachers also felt it greatly improved students' critical thinking in other subject areas. There was no clear consensus among the teachers whether one type of group composition was better than the other. Some felt that homogeneous grouping was better for

promoting basic reading comprehension but that heterogeneous grouping was better for encouraging participation. For example, one teacher wrote: "I am unsure which grouping is better. I think the students get a little more out of discussion in heterogeneous mixes, and I feel in most cases the discussions are a little better." On the other hand, a different teacher responded: "homogeneous grouping of the lower students brought out conversation that otherwise would not have happened." In sum, these data showed that the Quality Talk intervention was implemented with fidelity and that the small-group discussions promoted a context that afforded the kinds of social interactions thought to mediate the effect of grouping on reading achievement (Saleh et al., 2005).

Results

Descriptive Statistics

Descriptive statistics for all variables are provided in Table 1. Basic comprehension (i.e., multiple-choice assessment) scores increased from Baseline to Time 2 and then showed a slight decline at Time 3. Scores split by group composition showed similar patterns (see Table 2). On the other hand, high-level comprehension (i.e., written argumentation assessment) scores increased steadily over each consecutive time point. Means over the course of the intervention for the low-ability students grouped homogeneously showed growth very similar to that of their average- and high-ability peers from Baseline to Time 3 (see Table 3). By comparison, low-ability students grouped homogeneously performed somewhat poorly than their peers in terms of high-level comprehension, with lower mean scores than their peers at each time point. Importantly, however, strong gains were present for all students for all forms of comprehension, regardless of grouping over the yearlong intervention.

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

The correlation matrix (see Table 4) showed some correlations between consecutive time points, across outcome measures, but no clear pattern emerged. However, bivariate correlations are likely not informative given they do not account for text effects or other relevant covariates. Notably, group composition (i.e., homogeneous versus heterogeneous ability grouping) was not statistically significantly correlated with any other variables, but oral reading fluency at Baseline was statistically significantly correlated with basic and high-level comprehension scores at Baseline. Grade (i.e., 4^{th} or 5^{th}) was statistically significantly correlated with only basic comprehension and topic knowledge at Time 2 (i.e., 2 out of 10 student measures), suggesting that analyses could be conducted after collapsing across grades. Independent samples t-tests with grade as the grouping variable were statistically non-significant for both basic and high-level comprehension at each time point (all ts < 1.98, all ps > .05). Therefore, while we tested the predictive validity of grade in our multilevel models, we felt confident fitting multilevel models across grades.

Changes in Basic and High-Level Comprehension

Our first research question investigated whether students' basic and high-level comprehension changed over the course of the Quality Talk intervention, controlling for text and topic knowledge. Given that scores were nested in students, we utilized piecewise multilevel modeling to examine changes in comprehension over time. There were 62 students in our sample (i.e., level-2 units) with no missing data on level-2 predictors (i.e., ability grouping, oral reading fluency score, and grade). We had only six instances of missing data at level-1, which appeared to be missing completely at random (i.e., no clear missingness mechanism, particularly given the low percentage of missing data; Graham, 2009). All multilevel modeling was

conducted using the program HLM version 7.01 (Raudenbush, Bryk, & Congdon, 2010) with restricted maximum likelihood estimation.

Basic Comprehension. The intraclass correlation coefficient (ICC) for the basic comprehension outcome variable was .17, indicating that 17% of the variance in scores was due to variance between students (see Tables 5 and 6 for all basic comprehension model results). We utilized a model-building approach (Raudenbush & Bryk, 2002) to investigate level-1 and level-2 predictors. Specifically, we coded the time variable such that the first time point was zero, making the intercept the mean score at Baseline, adjusted for any other level-1 predictors. In our piecewise multilevel model, the variable Piece 1 represented the change in scores from Baseline to Time 2, and Piece 2 represented the change in scores from Time 2 to Time 3.

Our test of growth, without any level-2 predictors and modeling both piecewise variables as random effects, revealed that there was a statistically significant increase in basic comprehension scores from Baseline to Time 2, on average, but no such increase from Time 2 to Time 3 (see Table 5 for all models of basic comprehension using only level-1 predictors). Further, variance components for both piecewise variables were statistically significantly different than zero. These findings suggested that there was between-student variance in initial score and the two piecewise variables. Therefore, we explored modeling each variance using level-2 predictors.

Next, we included text as a level-1 predictor with fixed effects in the form of two dummy-coded variables with Text C as the comparison group (see Table 5, $Model_2$). Results indicated that text was a statistically significant predictor of basic comprehension scores; therefore, we kept these predictors in the model as controls. In our next model, we added the

topic knowledge score as a fixed effect, grand mean centered, which revealed that it was not a statistically significant predictor. Therefore, we dropped this predictor from our models.

Our examination of level-2 predictors began with the grade variable as a predictor of intercept and both piecewise variables. As expected, it was not a statistically significant predictor of any level-1 variable; therefore, we dropped this variable from our analyses[3]. Oral reading fluency was added as a grand mean centered level-2 predictor of intercept and both piecewise variables, and it statistically significantly moderated each level-1 effect except the change from Baseline to Time 2 (see Table 6, Model$_4$). These findings indicated that oral reading fluency positively correlated with scores at Baseline, and, on average, participants with higher reading fluency scores had a slightly more negative slope from Time 2 to Time 3 compared to their peers with lower reading fluency scores, who had a slightly more positive slope.

Our second research question involved an analysis of whether the growth trends identified in RQ1 were different depending upon whether students were randomly assigned to a homogeneous or heterogeneous group. We entered the grouping level-2 variable as an uncentered predictor of each level-1 random effect. Results indicated that group composition was not a statistically significant predictor of any level-1 parameters. Therefore, we removed statistically non-significant level-2 predictors from the models in a sequential manner until the best model fit was achieved. The final model for basic comprehension scores, which best fit the data given statistical significance of the various predictors and deviance score, was Model$_6$ (See Table 6).

---

[3] All models not reported in this manuscript are available upon request from the authors.

The final model for basic comprehension scores indicated that, on average, participants' score at Baseline was 4.03 out of a possible 9 points, for Text C. Scores on Texts A and B were lower, on average. As expected, oral reading fluency scores were positively correlated with basic comprehension score at Baseline. On average, participants increased their basic comprehension score from Baseline to Time 2, but there was no statistically significant change in average score from Time 2 to Time 3. Importantly, there was no evidence that group composition moderated growth in basic comprehension. The r-squared estimate for this model, using a formula that divides the change in within student variance between the null and final model by the within student variance of the null model, was .69, which converts into a Cohen's *d* value of 2.98, a large effect.

High-Level Comprehension. Our model-building process for the high-level comprehension outcome variable mirrored that for the basic comprehension variable (see Table 7). The ICC for high-level comprehension was less than one percent, indicating that the vast majority of variance in these scores was due to differences within students (e.g., time, topic knowledge, or text difficulty) as opposed to between students (e.g., group composition). Nonetheless, we proceeded to investigate all level-1 and level-2 predictors. Both Piece 1 (i.e., Baseline to Time 2) and Piece 2 (i.e., Time 2 to Time 3) were entered as uncentered level-1 predictors with random effects and both were statistically significant. This indicated that, on average, participants' high-level comprehension scores increased at each time point. Our investigation of level-1 control variables (i.e., text as an uncentered fixed effect and topic knowledge as a grand mean centered fixed effect) showed that topic knowledge was a statistically significant predictor, but text was not. Therefore, text was dropped from the model.

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Our investigation of grade as a level-2 predictor revealed that it did not statistically significantly predict any level-1 parameters; therefore, it was dropped from all models. Next, we added oral reading fluency as a grand mean centered level-2 predictor of all level-1 random effects. Oral reading fluency statistically significantly predicted variance in high-level comprehension scores only at Baseline (see Table 8, $Model_4$). Then we entered the group variable as a predictor of all level-1 random effects, but none were statistically significant (see Table 8, $Model_5$). Given that there was no rationale for grouping effects at Baseline, due to random assignment, we ran a model with grouping predicting only the Piece variables' random effects, and this model showed that grouping was a statistically significant moderator of the Piece 1 effect. We also ran a series of models, alternately adding and removing the reading fluency and grouping variables to each level-1 random effect, to determine the best model. Based upon examination of statistical significance and deviance scores, the best-fitting model for these data was the one with oral reading fluency predicting variance in the intercept and grouping predicting both Piece variables (see Table 8, $Model_6$). This model was our final model for high-level comprehension.

The final model for high-level comprehension indicated that participants' oral reading fluency scores were positively correlated with their Baseline high-level comprehension scores. On average, participants in both the homogeneous and heterogeneous groups increased their high-level comprehension from Baseline to Time 2, but this increase was statistically significantly greater for the heterogeneously grouped participants, compared to the homogeneously grouped students. Thus, in response to RQ2, heterogeneous grouping was superior to homogeneous grouping in terms of high-level comprehension from Baseline to Time 2. Neither group statistically significantly increased their performance from Time 2 to Time 3,

although there is some evidence that the homogeneously grouped students increased their scores, on average, more than the heterogeneous group students. The r-squared estimate for this model, using a formula that divides the change in within student variance between the null and final model by the within student variance of the null model, was .54, which converts into a Cohen's *d* value of 2.17, a large effect.

Summary of effects across basic and high-level comprehension analyses. The results of our analyses for RQ1 showed, on average and across both the basic and high-level comprehension outcome measures, students displayed a positive growth trajectory from Baseline to Time 2. These findings support the efficacy of Quality Talk as a way to use small-group discourse to bolster students' comprehension. The gains for low-ability students in terms of basic comprehension were particularly notable. In terms of RQ2, group composition did not have an effect upon basic comprehension, but it did influence high-level comprehension, with heterogeneous groups outperforming homogeneous groups, on average. As can be seen in Table 3, the low-ability students in homogeneous groups showed slightly less growth than other groups. The low-ability students displayed solid gains in basic comprehension during Quality Talk, but their gains in high-level comprehension were lower than those of their peers. To better understand these findings, we explored the group discourse that occurred during Quality Talk implementation via both quantitative (i.e., RQ3) and qualitative (i.e., RQ4) methods.

Quantitative Analysis of Fourth-Grade Discourse

Our third research question examined how homogeneous and heterogeneous ability groups differed with respect to *students'* discourse as indicators of critical-analytic thinking and argumentation as well as how the frequency of *teachers'* use of discourse moves differed between ability groups. To gather a richer understanding of the differences in group composition

with respect to changes in discourse, fourth-grade group discourse was analyzed at Time 1, Time 2, and Time 3 (i.e., intervention weeks 2, 10, and 19) for the homogeneous low-ability group, the homogeneous high-ability group, and one heterogeneous group. These particular groups were selected because all three were facilitated by the same teacher at all three times and to investigate how the homogeneous low-ability group's experience might differ from the two comparison groups: homogeneous high-ability and heterogeneous.

Student discourse. In our investigation of the effect of group composition, we examined changes in discourse indicators of students' critical-analytic thinking (i.e., the frequency of student-initiated authentic questions; Figure 1) and argumentation (i.e., the duration of students' elaborated explanations and exploratory talk events; Figure 2) over the length of the intervention and how they differed by group. At Time 1, students in all three groups already evidenced the capacity to ask authentic questions with only slight differences evidenced between the groups. While the duration of the argumentation responses was somewhat short at Time 1, this was likely due to the fact that the Quality Talk mini-lessons pertaining to argumentation had not yet been delivered; by Time 3, all groups evidenced increases in the duration of time spent engaging in argumentation responses. This suggests that students across all three groups engaged in the key aspects of Quality Talk in their discussions.

When examining the patterns of critical-analytic thinking and argumentation exhibited by students in the various groups, there were marked differences. Notably, discourse trends in the homogeneous high-ability group and the heterogeneous group were similar, whereas differences were apparent for the homogeneous low-ability group. The frequency of authentic questions (AQs) was initially high, but it decreased between Time 2 and Time 3 for students in the homogeneous high-ability group and the heterogeneous group. However, this decrease occurred

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

as students devoted more effort toward responding to, rather than asking, questions—a pattern evidenced by concomitant, and striking, increases in the duration of time students spent responding to the AQs with argumentation responses. In contrast, the discourse among students in the homogeneous low-ability group showed relatively consistent trends; indeed, the frequency of authentic questions showed a fairly flat slope and the duration of argumentation responses exhibited a positive but unremarkable slope. This suggests that these students may have acquired the ability to ask good questions early in the intervention, but subsequent argumentation patterns failed to mirror the growth observed in the other groups.

Teacher discourse. In addition to our investigation of the effect of group composition on students' critical-analytic thinking and argumentation, we also examined differences with respect to teachers' use of discourse moves (see Figure 3). For all three groups, the frequency of teacher discourse moves was the highest at Time 1 and over time became less frequent. Essentially, teachers initially provided a greater degree of scaffolding for students (e.g., prompting students to elaborate or marking good authentic questions), yet over time they faded their scaffolding by decreasing the use of these moves. However, the *degree* of decrease over time differed between the groups. At Time 2, the teachers used almost no discourse moves with the homogeneous high-ability group and the heterogeneous group, while the teachers continued to use discourse moves to facilitate the homogeneous low-ability group discussion.

Summary of coded discourse analyses. The quantitative discourse analyses provided a plausible explanation for why homogeneous low-ability participants, on average, did not experience the kind of growth in high-level comprehension that they did in basic comprehension. In essence, the homogeneous low-ability group exhibited the kinds of student discourse trends typical in the early stages of QT, when students use authentic questions to think *about* the text,

prompting basic comprehension. However, the homogeneous low-ability group did not display the growth in argumentation responses typical of students who have learned to think *around* and *with* the text, leading to high-level comprehension. In addition, at Time 2 teachers used more discourse moves to facilitate the homogeneous low-ability group compared to the other groups. This is further evidence that suggests that the students in the homogeneous low-ability group had different experiences from those in other groups. We investigated these differences in more detail via the qualitative analyses associated with RQ4.

Qualitative Analysis of Fourth-Grade Discourse

We thickened our quantitative discourse analyses in RQ3 by engaging in qualitative analyses for RQ4, focusing on the experience of the low-ability students, in both homogeneous and heterogeneous groups, compared to their high-ability and heterogeneously grouped peers. This allowed us to better understand why homogeneously grouped low-ability students evidenced slightly less gains in high-level comprehension compared to their peers. Our fourth research question was: In what ways did the experience of low-ability students differ across types of grouping, and differ from their high-ability peers?

To address this research question, three of the authors initially reviewed discussion videos for the selected groups at Time 1, Time 2, and Time 3. This included viewing the videos multiple times to get a sense of the discussions, to become familiar with the participants, and to become immersed in the data (Marshall & Rossman, 2015). Then, we began an in vivo, open-coding process by re-watching the videos and looking for patterns regarding differences between students' experience in homogeneous and heterogeneous grouping. Each researcher recorded notes and memos during open-coding.

After thoroughly combing through the data, the three coders met to discuss initial themes and categories. Similar themes were collapsed and seven general categories were agreed upon by the coders. With these reduced themes in mind, the coders compiled a thematic memo (Rossman & Rallis, 2003), inserting relevant situations and quotes that supported the themes. Once the coders felt the data were fully saturated and the thematic memo complete, the categories were discussed a final time and checked for internal convergence and external divergence (Guba, 1978). The seven categories were reduced to a final four relevant categories, which were then classified as belonging to one of two broad areas of interest: differences between high- and low-ability students' interactions irrespective of group composition (i.e., homogeneous or heterogeneous) and differences in how students interacted with each other within homogeneous ability groups versus heterogeneous ability groups. For this study, we focus on the findings for the first broad area of interest, which included three themes related to questions, responses, and engagement, as well as one finding from the second broad area of interest related to low-ability student engagement.

Question quality. Overall, low-ability students' questions were of lower quality than those of high-ability students. High-ability students asked more authentic questions prompting their peers for high-level thinking or connections to outside texts or content. These questions promoted richer discussion as there were multiple possibilities for students to consider when responding to the question. Questions asked by low-ability students often had a direct answer from the book (i.e., test questions), which did not foster as much discussion (see Table 9 for question examples). Also, the low-ability students asked vague questions that required teacher scaffolding or clarification (see Excerpt 1).

Excerpt 1 from Time 1 heterogeneous discussion:

TEACHER: Sequoia?

S14 (low): How high can a-what is it called?

S14 (low): Sequoia tree, how high can sequoia trees go?

S4 (average): It's actually a test question.

TEACHER: That's a test question.

S14 (low): Oh.

S4 (average): It says they can go over, um, 300 feet tall and 40 feet around.

TEACHER: Do you have a question that has more than one answer that we could answer? Like, what's the why question you have?

Each student wrote four questions in their Quality Talk literacy journal prior to their discussion. Students often used these questions to start the discussion or asked these questions during pauses to keep the conversation going. Prior to the discussion, teachers marked students' notebooks with a minus, a check, or a check plus to denote question quality. Not surprisingly, high-ability students had more check plusses than other students, across both homogeneous and heterogeneous groups. In most discussions, all of the students in the homogeneous high-ability group had at least one check plus. Low-ability students had fewer checks and check plusses, suggesting that writing authentic questions was a challenging task for these students. During Time 3, none of the students in the homogeneous low-ability group had a check plus. As might be anticipated, the differences in the kinds of questions students prepared for, and asked, during discussions appeared to parallel the quality of the responses provided within each discussion group.

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Response quality. Low-ability students across group composition types provided shallower responses than their high-ability peers and were less likely to explore and expand upon discussion topics with alternative opinions or challenge other group members using argumentation techniques. Rather, their discourse often more *about* understanding the text as seen in Excerpt 1. In the homogeneous high-ability group, students tended to focus on reasoning *around* and *with* the text, and there were long periods of dialogic exchange between students without the teacher being involved. As was evident in the analysis of teachers' use of discourse moves from RQ3, teachers released responsibility and interpretive authority to those in the homogeneous high-ability group fairly early in the Quality Talk intervention (see Excerpt 2).

Excerpt 2 from Time 2 homogeneous high-ability discussion:

S1 (high): What if Luke only named one cloud instead of seven?

S10 (high): Then there would be only one cloud name, and we would probably have a hard time figuring out which cloud is going to bring rain.

S12 (high): And storms.

S1 (high): I don't--I dis--I don't--

S21 (high): Well, what if somebody else continued his thing--

S1 (high): I disagree with that, (S10).

S21 (high): --and kept naming the clouds?

S1 (high): I disagree with that, (S10), because I think the friend--if he only named one cloud, I think, um, his rival would win because he probably had more clouds than--named than him.

S10 (high): Well, but if the rival won, everyone would be, like, what's the name of the cloud again? But he'd still have one cloud. Um--he'd still have one cloud, and they'd

be, like, oh that's a cloud. And they're, like, and then maybe they'd base more words off of that cloud--

S1 (high): Well, actually--

S10 (high): --if he only named one.

S1 (high): --I think that our clouds today would not be named. Our clouds today. They would be mixed up. So we probably wouldn't be here.

S10 (high): Cumulonimbus just might be a stratus cloud.

S1 (high): Yeah. You never know.

Low-ability students provided more surface-level responses than their peers. These students usually agreed with other students without asking them to provide reasoning or evidence. The teacher had to guide them back on topic more often than high-ability students. After getting a sense of the text (i.e., basic comprehension or knowing about the text), these students tended to turn toward funny ponderings, opinions, or personal stories (i.e., *around* the text), making connections to their own experiences. While relating to the text through personal stories can be an effective way to engage with the text, it was less likely to lead to high-level comprehension without teacher facilitation. Teachers often attempted to push low-ability students toward deeper, richer understandings of the text with discourse moves. Relatedly, as can be seen in the excerpt below, teachers' may have found it particularly challenging to release responsibility to the students and allow for their interpretive authority in the homogeneous group containing all low-ability students (see Excerpt 3).

Excerpt 3 from Time 1 homogeneous low-ability discussion:

TEACHER: --several days. What would you take with you for several days?

S7 (low): Lots of water.

S13 (low): I know.

TEACHER: Lots of water? OK. Go ahead.

S13 (low): Um, I would bring a [inaudible]. A life supply of gum.

TEACHER: Gum? Why?

S13 (low): Because--

TEACHER: (inaudible)

S13 (low): Because it's sticky, and it can help me climb.

S20 (low): You would probably--I would probably bring this. A life supply of chocolate candies.

TEACHER: You don't need a life supply for a few days.

Overall, the discourse from high-ability students was rich with elaborated explanations and exploratory talk with and around the text, while the discourse from low-ability students frequently deviated away from the text, leading to overall shallower engagement with text over the course of the discussion, as compared to the high-ability students. Importantly, however, the basic comprehension gains of low-ability grouped students exceeded any other type of grouping. It may be that this surface-level discourse scaffolded their basic understandings of the text—understandings requisite for future high-level comprehension.

Engagement with text. During discussions, low-ability students typically did not engage with the text unless prompted by the teacher or another student. In the Time 1 discussion for the homogeneous low-ability group, the students rarely referred to the text. When discussing the question: "How long will it take to climb El Capitan?" students made guesses from days to years. The teacher had to refer students to the book where it said: "It can take anywhere from several hours to several days to scale this rock." During the Time 2 discussion, the teacher explicitly

told students in the homogeneous low-ability group to refer to their book, as some students did not have it open: "Well, it says that--on page 327--why don't we have everyone open their book, OK? So you have it… if you need to refer to it, you have it." High-ability students often referred to the text without prompting.

In the homogeneous low-ability group, facts about text were not discussed as much, but when they were, errors were often not corrected by other students in the group, particularly early in the intervention. In heterogeneous groups, average- and high-ability students would often clarify errors about the text; in essence, these students were enacting interpretive authority. There were times in heterogeneous groups when some students would look through the book to find a piece of information relevant to the conversation and low-ability students would disengage rather than try to find an answer. High-ability students would refer to the text in general as well as point to specific page numbers (see Excerpt 4).

Excerpt 4 from Time 1 heterogeneous discussion:

TEACHER: Why would you be afraid to be lost?

S5 (average): Because you're like alone and you're like without anyone else.

S17 (average): And there's no hot dog stands around.

[S2 starts flipping through her book]

S5 (average): There could be like dangerous things.

S2 (high): [reading number from text] It is 1,170 square miles and your parents could be on the other side of the park.

Engagement in the discussion. One of the main differences across homogeneous and heterogeneous ability group types was students' engagement in the discussion. For example, low-ability students in the homogeneous group were most animated when talking about their

own personal experiences or when listening to others' stories loosely related to the text (see Excerpt 5).

Excerpt 5 from Time 3 homogeneous low-ability discussion:

S9 (low): And it also--in *Planes: Fire & Rescue*--they also tried to rescue the animal. Remember the deer that was way behind? In the--

S7 (low): And then they go back.

S9 (low): --yeah. And then, um, the, um, helicopter guy--

S7 (low): Yeah.

S9 (low): --picked them up with his hook, and he put them in front of the line.

TEACHER: All right, kids.

S7 (low): That was funny.

Low-ability students in the heterogeneous groups, however, were more reserved and hesitant than their low-ability peers in the homogeneous group. These students also asked fewer questions, and their responses were infrequent and brief. When they did participate, these low-ability students would often talk softly or only respond when directly asked. Exchanges between low-ability and high-ability students were somewhat hegemonic with high-ability students correcting, challenging, or even talking over low-ability group members. High-ability students, on the other hand, were animated and comfortable in discussions, regardless of the type of group composition. However, despite being engaged in discussion when in homogeneous groups, low-ability students did not routinely enact the kinds of argumentation supportive of high-level comprehension. While their low-ability peers in heterogeneous groups did not engage deeply with the discourse either, low-ability students in heterogeneous groups were at least present as high-level argumentation occurred among their high- and average-ability peers.

Summary. In sum, low- and high-ability students engaged in the discussions differently, which seemed to play an interesting role within the homogeneous low-ability group. High-ability students asked deeper, more relevant questions than their low-ability students. High-ability students also produced more in-depth responses, perhaps because of their greater engagement with the text, compared to low-ability students. Students in the homogeneous low-ability group focused on talking around the text, rather than engaging with the text itself or the topic of that text. In essence, the text provided a jumping off point from which the low-ability students would move to discuss personal stories or anecdotes of interest. This type of discourse can promote basic comprehension but not necessarily ensure high-level comprehension. Low-ability students in heterogeneous groups were present, if not engaged, during discourse that promotes high-level comprehension. Unfortunately, their engagement was minimal and somewhat meek compared to other group members. The results of RQ4 provide a plausible explanation for why students in the homogeneous low-ability group slightly underperformed their peers in terms of high-level, but not basic, comprehension. Their discourse did not often contain the questions, responses, or engagement necessary for critical-analytic thinking, a key predictor of high-level comprehension skills (Soter et al., 2008).

Discussion

Small-group discussions are a prominent and useful instructional practice, compared to whole-class instruction (Lou et al., 1996). However, the literature on the effects of homogeneous versus heterogeneous ability grouping has been inconclusive with evidence for and against each form of group composition (Murphy et al., 2009). Further, there is some indication that the quality of social interactions within groups mediates the effect of grouping on students' reading achievement (Saleh et al., 2005). Thus, there is a need for research examining the effects of

group composition on both the quality of small-group discussion interactions as well as reading achievement. Such research must be done within classroom contexts that promote the kinds of social interactions necessary for high-level comprehension of text (Murphy et al., 2009).

The purpose of this study was to investigate how different ways of composing small-group discussions (i.e., homogeneous versus heterogeneous ability grouping) influenced students' discourse and subsequent basic and high-level comprehension. Using oral reading fluency scores at Baseline, we created matched pairs of students with similar ability, and then randomly assigned each student to either a homogeneous or heterogeneous ability group for discussion. Our results showed that oral reading fluency was a statistically significant predictor of students' high-level comprehension scores at Baseline, which further supports our rationale for employing oral reading fluency as the grouping variable. All students then engaged in Quality Talk, a theoretically- and empirically-supported intervention for using small-group discussions to promote high-level comprehension of text. Over the course of an academic year, all students learned how to engage in productive discourse including how to ask authentic questions and engage in argumentation to appropriately challenge each other and explore and understand texts. Our measures of basic and high-level comprehension revealed that, on average, students exhibited practically and statistically significant gains on both basic and high-level comprehension over the duration of the Quality Talk intervention. However, whereas group composition did not have a differential effect on basic comprehension, it did in terms of high-level comprehension. Heterogeneously grouped students exhibited larger gains from Baseline to Time 2, on average, than homogeneously grouped students. Such a finding may seem to contradict previous meta-analytic findings which showed an advantage of homogeneous over heterogeneous grouping in reading (Lou et al., 1996). However, this may be because few, if

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

any, of the studies controlled for fluency or examined students' *high-level* comprehension. Thus, our study, which assesses both basic and high-level comprehension, provides a valuable addition to the extant literature. For researchers and teachers interested in high-level comprehension, in particular, there is evidence to support the use of heterogeneous, small groups in classroom discussions.

Our analyses of group discourse, using frequency counts of authentic questions and the duration of argumentation responses as indicators of high-level comprehension, as well as qualitative analysis of the discourse itself, also revealed important differences related to the group composition. Low-ability students in the homogeneous groups did not display the kinds of discourse likely to foster high-level comprehension. This finding aligns with the previous research that supported the use of heterogeneous grouping as a mechanism to promote low-ability students' learning outcome and social interaction in group processes. However, low-ability students were less engaged in heterogeneous groups, and were rarely granted interpretative authority by other groups members. The long-term consequences of these types of exchanges for low-ability students are unclear. In sum, our findings indicated heterogeneous grouping was more beneficial than homogeneous grouping for high-level comprehension, on average, but there remains more work to be done to involve low-ability students in discourse aimed at high-level comprehension.

Limitations

Our findings are contextualized in several ways. First, our experimental manipulation involved the type of group composition but not the Quality Talk intervention itself. All students received the Quality Talk intervention, therefore we cannot make causal claims about its efficacy

compared to other literacy interventions or activities. However, the gains in basic and high-level comprehension in this study meet or exceed those of many other literacy interventions (Lou et al., 1996; Murphy et al., 2009). Second, despite being a relatively large sample for a yearlong experimental study (Murphy, 2015), this study was underpowered to examine interactions between group composition and measured oral reading fluency. Our analyses show no evidence of such interactions, but we were unable to statistically test them. Finally, our findings are situated within a particular school and sample context. Generalization of these findings to a different context must be done with care until additional research can be done on Quality Talk and group composition in other contexts.

Future Directions and Implications

The promising results of this study, particularly in terms of the efficacy and fidelity of Quality Talk implementation, the differences in high-level comprehension by group composition, as well as the interesting findings regarding the students' experience in the discussions, suggest several future directions for research. To begin, quasi-experimental and randomized control trials of the Quality Talk intervention are needed to make causal claims about its efficacy. Likewise, studies with larger samples, including control groups, could be used to further explore interactions between group composition and ability level as well as to establish causal mechanisms and the relevance of possible mediators such as social interactions within discourse. As with any study, replications of this study in different and diverse contexts would create a body of literature on not only the active ingredients of Quality Talk but also the factors that might moderate their efficacy and applicability (Greene, 2015). Finally, additional qualitative research involving analysis of student discourse could shed light upon the differential

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

engagement of low-ability students in small-group discourse and perhaps identify promising directions for fostering these students' engagement and critical-analytic stance.

There are two important implications from our findings. First, this study adds to the growing corpus of research demonstrating the effectiveness of Quality Talk as a way to promote both basic and high-level comprehension. On average, all students benefitted from Quality Talk, and growth on our comprehension measures exceeded typical yearlong gains in achievement (e.g., Cohen's d = .41 for 4th grade students' average growth in basic comprehension; Scammacca, Fall, & Roberts, 2014). Importantly, the results reported in this paper also compliment those found in previous studies of Quality Talk (Firetto, Murphy, Greene, Li, Wei, Montalbano, Hendrick, & Croninger; Li et al., 2016; Murphy, Greene, Firetto, Hendrick, Montalbano, Li, & Wei, 2016). For example, in one previous study (Murphy et al., 2016), students participating in Quality Talk evidenced statistically significant growth in their written argumentation scores (i.e., high-level comprehension) with a small to medium effect (i.e., Cohen's d = .35), which was greater than the effect sizes reported for Collaborative Reasoning (i.e., ES = 0.26) and Philosophy for Children (i.e., ES = 0.21). While the effect sizes reported in our Results (i.e., Cohen's $d > 2$) are not directly comparable due to the predictors included within the multilevel modeling, the accumulation and triangulation of converging evidence is certainly noteworthy as our analysis accounts for variance (e.g., oral reading fluency) not accounted for in studies of other discussion approaches.

Second, our qualitative student discourse data point to numerous implications for the implementation of small-group discourse with struggling readers. While additional research is necessary before firm guidelines can be forwarded for practice, based on the findings reported herein, we can offer several tentative suggestions for teachers to consider as they employ small-

group discussions in their classrooms. Teachers must be mindful of the balance between fostering students' engagement and keeping the students focused on discourse about, around, and with the text. It may be that teachers should employ targeted discourse moves to better support the engagement of low-ability students in heterogeneous groups. Further, while personal connections to the text can be facilitated through personal anecdotes, those connections should be used to guide students back to critical-analytic thinking about the text and topic, particularly for low-ability students in homogeneous groups. Additionally, teachers' goals and expectations for small-group discussions should guide their decision to compose the groups homogeneously or heterogeneously. For example, if teachers desire to focus on enhancing students' basic comprehension or if they desire to support students' engagement in the discussion, they may find that grouping the students homogeneously is more advantageous for low-ability students. Alternatively, teachers should employ heterogeneous ability grouping if their focus is on building students' high-level comprehension of the text. Finally, it may be necessary for teachers to also employ discourse moves that foster low-ability students' interpretive authority in heterogeneous groups. Overall, our quantitative and qualitative findings suggest that Quality Talk is a promising intervention, and future studies will focus on accumulating the necessary research to test better mechanisms for encouraging low-ability groups' engagement and allow us to forward more concrete recommendations.

Conclusion

In sum, our research showed that students displayed atypical levels of growth in basic and high-level comprehension over the course of the Quality Talk intervention. Such findings demonstrate the promise of theoretically- and empirically-derived literacy interventions that include explicit instruction of productive discourse and support for teacher scaffolding and

fading. Further, this study qualifies and expands upon the within-class ability grouping literature by showing that basic comprehension can be achieved by students regardless of group composition but that heterogeneous grouping is more likely to lead to high-level comprehension than homogeneous grouping. That being said, additional research is needed to understand how to foster the kinds of accelerated growth needed to help lower-ability students reach the levels of high-level comprehension displayed by their higher-ability peers. Our analyses of student discourse point to several promising directions for future research and practice, including ways to help students harness their personal connections to texts in ways that promote a critical-analytic stance toward the text. Given the pervasive use of small-group instructional practices in classrooms, particularly those focused on text-based comprehension, it is imperative that researchers continue to explore the utility and feasibility of approaches, like Quality Talk, for enhancing the literacy skills of all students.

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

References

Alexander, P. A., Kulikowich, J. M., & Schulze, S. K. (1994). How subject-matter knowledge affects recall and interest. *American Educational Research Journal, 31*, 313–337.

Allington, R. L. (1980). Poor readers don't get to read much in reading groups. *Language Arts, 57*, 872–877.

Azmitia, M. (1988). Peer interaction and problem solving: When are two heads better than one? *Child Development, 59*, 87–96.

Barr, R. (1975). Influence of reading materials on responses to printed words. *Journal of Reading Behavior, 7*, 123–135.

Bråten, I., Britt, M. A., Strømsø, H. I., & Rouet, J. F. (2011). The role of epistemic beliefs in the comprehension of multiple expository texts: Toward an integrated model. *Educational Psychologist, 46*(1), 48–70.

Coldiron, J. R., Braddock, J. H., & McPartland, J. M. (1987, April). *A description of school structures and classroom practices in elementary, middle, and secondary schools.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Common Core English Language Arts Standards. (2011). Retrieved from http://www.corestandards.org/ELA-Literacy

Eder, D., & Felmlee, D. (1984). Development of attention norms in ability groups. In P. L. Peterson, L. C. Wilkinson, & M. Hallinan (Eds.), *The social context of instruction: Group organization as group processes* (pp. 189–208). New York: Academic Press.

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Firetto, C. M., Murphy, P. K., Greene, J. A., Li, M., Wei, L., Montalbano, C., Hendrick, B., & Croninger, R. M. V. (2016, April). *Using Quality Talk to foster transfer of students' critical-analytic discussions to their argumentative writing.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Karns, K. (1998). High-achieving students' interactions and performance on complex mathematical tasks as a function of homogeneous and heterogeneous pairings. *American Educational Research Journal, 35*, 227–267.

Fuchs, L., Fuchs, D., Hosp, M., & Jenkins, J. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*(3), 239–256.

Goatley, V. J., Brock, C. H., & Raphael, T. E. (1995). Diverse learners participating in regular education Book Clubs. *Reading Research Quarterly, 30*(3), 352–380.

Goffreda, C., & Diperna, J. (2010). An empirical review of psychometric evidence for the dynamic indicators of basic early literacy skills. *School Psychology Review, 30*(3), 463–483.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549–576.

Greene, J. A. (2015). Serious challenges require serious scholarship: Integrating implementation science into the scholarly discourse. *Contemporary Educational Psychology, 40*, 112–120.

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51*, 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Greene, J. A., Sandoval, W. A., & Bråten, I. (2016). Introduction to epistemic cognition. In J. A. Greene, W. A. Sandoval, & I. Bråten (Eds.), *Handbook of epistemic cognition* (pp. 1–15). New York: Routledge.

Guba, E. G. (1978). *Toward a methodology of naturalistic inquiry in education evaluation* (monograph 8). Los Angeles: UCLA Center for the Study of Evaluation.

Hadwin, A. F., Jarvela, S., & Miller, M. (2011). Self-regulated, co-regulated, and socially shared regulation of learning. In B. Zimmerman & D. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 65–84). New York, NY: Routledge.

Johnson, E., Jenkins, J., Petscher, Y., & Catts, H. (2009). How can we improve the accuracy of screening instruments? *Learning Disabilities Research & Practice, 24*(4), 174–185.

Johnson, D. W., Johnson, R. T., & Stanne, M. B. (2000). *Cooperative learning methods: A meta-analysis*. Minneapolis, MN: University of Minnesota, Cooperative Learning Center. Retrieved from https://www.researchgate.net/profile/David_Johnson50/publication/220040324_Cooperative_Learning_Methods_A_Meta-Analysis/links/00b4952b39d258145c000000.pdf

Johnson, D. W., Skon, L., & Johnson, R. (1980). Effects of cooperative, competitive, and individualistic conditions on children's problem-solving performance. *American Educational Research Journal, 17*, 83–93.

Kendeou, P., & van den Broek, P. (2007). Interactions between prior knowledge and text structure during comprehension of scientific texts. *Memory and Cognition, 35*, 1567–1577.

Kintsch, W. (1998). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review, 95*, 163–182.

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Kulik, J. A. (1992). *An analysis of research on ability grouping: Historical and contemporary perspectives.* Storrs, CT: University of Connecticut, National Research Center on the Gifted and Talented. Retrieved from ERIC database. (ED350777)

Kulik, J. A., & Kulik, C-L. C. (1987). Effects of ability grouping on student achievement. *Equity and Excellence, 23*, 22–30.

Lou, Y., Abrami, P. C., & Spence, J. C. (2000). Effects of within-class grouping on student achievement: An exploratory model. *Journal of Educational Research, 94*(2), 101–112.

Lou, Y., Abrami, P. C., Spence, J. C., Poulson, C., Chambers, B., & d'Apollonia, S. (1996). Within-class grouping: A meta-analysis. *Review of Educational Research, 66*, 423–458.

Marshall, C., & Rossman, G. B. (2015). *Designing qualitative research* (6th ed.). Thousand Oaks, CA: Sage Publications.

Metz, M. (1978). *Classrooms and corridors: The crisis of authority in desegregated secondary schools*. Berkeley, CA: University of California Press.

McKeown, M. G., Beck, I. L., & Blake, R. G. K. (2009). Rethinking reading comprehension instruction: A comparison of instruction for strategies and content approaches. *Reading Research Quarterly, 44*(3), 218–252.

McNamara, D. S., Kintsch, E., Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*, 1–43.

Murphy, P. K. (2007). The eye of the beholder: The interplay of social and cognitive components in change. *Educational Psychologist, 42,* 41–53.

Murphy, P. K. (2015). Marking the way: School-based interventions that "work." *Contemporary Educational Psychology, 40*, 1–4.

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Murphy, P. K., Firetto, C. M., Greene, J. A., & Butler, A. M. (2017). Analyzing the talk in Quality Talk discussions: A coding manual. doi.org/10.18113/S1XW64

Murphy, P. K., Greene, J. A., Firetto, C. M., Hendrick, B., Montalbano, C., Li, M., & Wei, L. (2016, April). *Enhancing students' comprehension and critical-analytic thinking through Quality Talk Discussions.* Poster presented at the annual meeting of the American Educational Research Association, Washington, DC.

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

National Assessment Governing Board. (2015). *Reading framework for the 2015 National Assessment of Educational Progress.* Washington, DC: U.S. Department of Education. Retrieved from: https://www.nagb.org/content/nagb/assets/documents/publications/frameworks/reading/2015-reading-framework.pdf

Piaget, J. (1932). *The language and thought of the child* (2nd ed.). London: Routledge & Kegan Paul.

Popham, W. J. (2006). *Assessment for educational leaders.* Upper Saddle River, NJ: Pearson Education, Inc.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Thousand Oaks, CA: Sage Publications.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. T. (2010). *Hierarchical linear modeling with the HLM7 programs.* Chicago, IL: Scientific Software International.

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Rosenbaum, J. E. (1980). Social implications of educational grouping. *Review of Research in Education, 8,* 361–401.

Rossman, G. B., & Rallis, S. F. (2003). *Learning in the field: An introduction to qualitative research* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Saleh, M., Lazonder, A. W., & De Jong, T. D. (2005). Effects of within-class ability grouping on social interaction, achievement, and motivation. *Instructional Science, 33*(2), 105–119.

Saleh, M., Lazonder, A. W., & De Jong, T. D. (2007). Structuring collaboration in mixed-ability groups to promote verbal interaction, learning, and motivation of average-ability students. *Contemporary Educational Psychology, 32*(3), 314–331.

Scammacca, N. K., Fall, A-M., & Roberts, G. (2014). Benchmarks for expected annual academic growth for students in the bottom quartile of the normative distribution. *Journal of Research on Educational Effectiveness, 8*(3), 366–379.

Shinn, M. R., Good, R., Knutson, N., Tilly, W., & Collins, V. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review, 21,* 459–479.

Shinn, M. R., & Shinn, M. M. (2002). *AIMSweb training workbook: Administration and scoring of reading Maze for use in general outcome measurement.* Eden Prairie, MN: Edformation.

Singer, J. D., & Willet, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York: Oxford University Press.

Slavin, R. E. (1987). Ability grouping and student achievement in elementary schools: A best-evidence synthesis. *Review of Educational Research, 57,* 293–336.

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Slavin, R. E. (1991). Synthesis of research on cooperative learning. *Educational Leadership, 48,* 71–82.

Slavin, R. E., (2011). Instruction based on cooperative learning. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of Research on Learning and Instruction* (pp. 344–360). New York: Routledge.

Soter, A. O., Wilkinson, I. A. G., Murphy, P. K., Rudge, L., & Reninger, K. (2006). *Analyzing the discourse of discussion: Coding manual.* Unpublished manuscript, The Ohio State University and The Pennsylvania State University.

Soter, A. O., Wilkinson, I. A. G., Murphy, P. K., Rudge, L., Reninger, K., & Edwards, M. (2008). What the discourse tells us: Talk and indicators of high-level comprehension. *International Journal Educational Research, 47,* 372–391.

Torgesen, J., Myers, D., Schirm, A., Stuart, E., Vartivarian, S., Mansfield, W., … & Haan, C. (2006). *National assessment of Title I interim report to congress—Volume II: Closing the reading gap: First year findings from a randomized trial of four reading interventions for striving readers.* Washington, DC: Institute of Education Science. Retrieved from http://www.ed.gov/rschstat/eval/disadv/title1interimreport/index.html

Tudge, J. (1989). When collaboration leads to regression: Some negative consequences of socio-cognitive conflict. *European Journal of Social Psychology, 19*(2), 123–138.

Vygotsky, L. S. (1978). *Mind in society: The development of higher mental processes.* Cambridge, MA: Harvard University Press.

Webb, N. M. (1980). A process-outcome analysis of learning in group and individual settings. *Educational Psychologist, 15,* 69–83.

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Webb, N. M. (1991). Task-related verbal interaction and mathematics learning in small groups. *Journal for Research in Mathematics Education, 22,* 366–389.

Webb, N. M., Nemer, K. M., Chizhik, A. W., & Sugrue, B. (1998). Equity issues in collaborative group assessment: Group composition and performance. *American Educational Research Journal, 35,* 607–651.

Webb, N. M., & Palinscar, A. S. (1996). Group processes in the classroom. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 841–873). New York: MacMillan.

Wilkinson, I. A., & Fung, I. Y. (2002). Small-group composition and peer effects. *International Journal of Educational Research*, *37*(5), 425–447.

Wilkinson, I. A. G., Soter, A. O., & Murphy, P. K. (2010). Developing a model of Quality Talk about literary text. In M. G. McKeown & L. Kucan (Eds.), *Bringing reading research to life* (pp. 142–169). NY: Guilford Press.

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Table 1

Descriptive Statistics of Individual Student Outcome Measures, Grouping, Oral Reading Fluency, and Grade

Variable	N	Mean (SD)	Skewness (SE)	Kurtosis (SE)
Basic Comprehension				
Baseline	62	3.55 (1.46)	-0.71 (0.30)	-0.55 (0.60)
Time 2	62	4.22 (1.03)	-1.40 (0.30)	1.62 (0.60)
Time 3	61	3.98 (1.22)	-1.11 (0.31)	0.74 (0.60)
High-Level Comprehension				
Baseline	60	3.02 (1.85)	0.36 (0.31)	-0.50 (0.61)
Time 2	62	4.00 (1.84)	0.15 (0.30)	-0.56 (0.60)
Time 3	61	4.98 (2.17)	1.16 (0.31)	2.30 (0.60)
Topic Knowledge				
Baseline	61	1.39 (1.64)	1.58 (0.31)	2.46 (0.60)
Time 2	62	1.36 (1.27)	0.49 (0.30)	-1.00 (0.60)
Time 3	61	1.90 (1.69)	1.42 (0.31)	2.52 (0.60)
Grouping[a]	62	0.50 (0.50)	0.00 (0.30)	-2.07 (0.60)
Oral Reading Fluency	62	148.02 (35.69)	-0.66 (0.30)	-0.29 (0.60)
Grade[b]	62	0.55 (0.50)	-0.20 (0.30)	-2.03 (0.60)

[a] Group composition was coded with zero indicating heterogeneous grouping and one indicating homogeneous grouping.
[b] Grade was coded with zero indicating Grade 4 and one Grade 5.

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Table 2

Descriptive Statistics of Individual Student Outcome Measures, by Group Composition

	Heterogeneous Grouping				Homogeneous Grouping			
Variable	N	Mean (SD)	Skewness (SE)	Kurtosis (SE)	N	Mean (SD)	Skewness (SE)	Kurtosis (SE)
Basic Comprehension								
Baseline	31	3.61 (1.38)	-0.94 (0.42)	0.26 (0.82)	31	3.48 (1.55)	-0.55 (0.42)	-1.00 (0.82)
Time 2	31	4.13 (1.09)	-1.60 (0.42)	2.68 (0.82)	31	4.32 (0.98)	-1.17 (0.42)	0.11 (0.82)
Time 3	31	4.03 (1.28)	-.129 (0.42)	1.60 (0.82)	30	3.93 (1.17)	-0.96 (0.43)	-0.02 (0.83)
High-Level Comprehension								
Baseline	30	2.87 (1.81)	0.40 (0.43)	-0.66 (0.83)	30	3.17 (1.90)	0.33 (0.43)	-0.26 (0.83)
Time 2	31	4.45 (1.84)	0.13 (0.42)	-0.86 (0.82)	31	3.55 (1.75)	0.11 (0.43)	-0.33 (0.82)
Time 3	31	5.03 (2.26)	1.74 (0.42)	4.09 (0.82)	30	4.93 (2.12)	0.49 (0.43)	0.38 (0.83)

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Table 3

Descriptive Statistics of Individual Student Outcome Measures, Homogeneous Groups by Ability

		Mean (SD)				
Variable	N	Low-Ability	N	Average-Ability	N	High-Ability
Basic Comprehension						
Baseline	10	2.80 (1.81)	11	3.64 (1.29)	9	4.33 (1.00)
Time 2	10	4.10 (1.10)	11	4.55 (0.82)	9	4.56 (0.73)
Time 3	9	4.56 (0.73)	11	3.45 (1.29)	9	3.78 (1.20)
High-Level Comprehension						
Baseline	10	2.60 (1.65)	10	3.10 (2.51)	9	4.00 (1.23)
Time 2	10	2.80 (1.62)	11	3.64 (2.06)	9	4.22 (1.39)
Time 3	10	4.40 (2.88)	10	5.30 (1.70)	9	5.22 (1.72)

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Table 4

Correlation Matrix of Independent and Dependent Variables

Level-1 Correlations	1	2	3	4	5	6	7	8	9	10	11	12
1. Basic Comprehension: Baseline	1											
2. Basic Comprehension: Time 2	.52**	1										
3. Basic Comprehension: Time 3	.09	.00	1									
4. High-Level Comprehension: Baseline	.28*	.32*	.14	1								
5. High-Level Comprehension: Time 2	.20	.39**	-.04	.17	1							
6. High-Level Comprehension: Time 3	.04	.18	.13	-.14	.25*	1						
7. Topic Knowledge: Baseline	.15	.41	-.12	.19	.19	.06	1					
8. Topic Knowledge: Time 2	.17	.20	-.19	.21	.09	-.03	.23	1				
9. Topic Knowledge: Time 3	.26*	.27*	.15	.32*	.16	.23	.16	.23	1			
10. Grouping[a]	-.05	.10	-.04	.08	-.25	-.02	.22	.13	.12	1		
11. Oral Reading Fluency	.37**	.48**	-.19	.40**	.31	.10	.27*	.37**	.33**	.01	1	
12. Grade	.12	.30*	.07	.246	.23	-.05	.061	.31*	.02	.00	.21	1

Note. * $p < .05$, ** $p < .01$

[a] Dichotomous variable indicating group composition, either heterogeneous, coded as zero, or homogeneous, coded as one.

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Table 5

Multilevel Models for Basic Comprehension Outcome Variable Using Only Level-1 Predictors

	$Model_0$		$Model_1$		$Model_2$		$Model_3$	
Variable	Est.	*SE*	Est.	*SE*	Est.	*SE*	Est.	*SE*
Fixed Effects								
Intercept	3.92***	0.11	3.55***	0.19	4.02***	0.19	4.09***	0.18
Piece 1			0.68***	0.16	0.71***	0.14	0.67***	0.13
Piece 2			-0.24	0.20	-0.27	0.21	-0.29	0.21
Text A					-0.85***	0.16	-0.82***	0.16
Text B					-0.62***	0.16	-0.67***	0.15
Topic Knowledge							0.05	0.05
Random Effects								
Intercept	0.28**		1.52***		1.30***		1.14***	
Piece 1			0.43*		0.29*		0.22	
Piece 2			1.34***		1.85***		1.87***	
Within Student	1.34		0.60		0.43		0.41	
Deviance	610.90		582.50		562.49		551.20	
Number of estimated parameters	2		7		7		7	

Note. Text A and Text B were dummy-coded variables, with Text C as the comparison group for each. Topic Knowledge variable was entered as grand-mean centered.
*$p < .05$, **$p < .01$, ***$p < .001$.

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Table 6

Multilevel Models for Basic Comprehension Outcome Variable Including Level-1 and Level-2 Predictors

	$Model_4$		$Model_5$		$Model_6$	
Variable	Est.	*SE*	Est.	*SE*	Est.	*SE*
Fixed Effects						
Intercept	4.03***	0.17	4.12***	0.23	4.03***	0.17
Reading Fluency	0.02***	<0.01	0.01***	<0.01	0.01***	<0.01
Grouping			-0.16	0.31		
Piece 1	0.71***	0.26	0.54**	0.19	0.71***	0.14
Reading Fluency	<0.01	<0.01				
Grouping			0.35	0.27		
Piece 2	-0.26	0.19	-0.12	0.26	-0.26	0.19
Reading Fluency	-0.02***	0.01	-0.02**	<0.01	-0.02***	<0.01
Grouping			-0.27	0.37		
Text A	-0.86***	0.15	-0.86***	0.15	-0.87***	0.15
Text B	-0.64***	0.15	-0.66***	0.15	-0.64***	0.15
Random Effects						
Intercept	1.04***		1.05***		1.03***	
Piece 1	0.57*		0.31*		0.31*	
Piece 2	1.27***		1.30***		1.27***	
Within Student	0.42		0.44		0.42	
Deviance	563.70		554.947		556.26	
Number of estimated parameters	7		7		7	

Note. Text A and Text B were dummy-coded variables, with Text C as the comparison group for each. Grouping was a dummy-coded variable with heterogeneous grouping coded as zero and homogeneous grouping coded as one.
Topic Knowledge variable not shown because it was not a statistically significant level-1 predictor.
Reading Fluency variable was entered as grand-mean centered. Grouping variable was entered as uncentered.
*$p < .05$, **$p < .01$, ***$p < .001$.

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Table 7

Multilevel Models for High-Level Comprehension Outcome Variable Using Only Level-1 Predictors

	$Model_0$		$Model_1$		$Model_2$		$Model_3$	
Variable	Est.	*SE*	Est.	*SE*	Est.	*SE*	Est.	*SE*
Fixed Effects								
Intercept	4.01***	0.16	3.01***	0.24	3.10***	0.30	3.05***	0.23
Piece 1			0.99**	0.30	1.00**	0.31	0.99**	0.31
Piece 2			0.99**	0.32	0.98**	0.32	0.86**	0.32
Text A					-0.12	0.32		
Text B					-0.17	0.32		
Topic Knowledge							0.22*	0.10
Random Effects								
Intercept	0.02		1.27*		1.27*		1.13*	
Piece 1			1.37*		1.31		1.64*	
Piece 2			1.84*		1.89*		1.72*	
Within Student	4.41		2.13		2.17		2.16	
	792.77		755.21		756.11		750.99	
Number of estimated parameters	2		7		7		7	

Note. Text A and Text B were dummy-coded variables, with Text C as the comparison group for each. Topic Knowledge was entered as grand-mean centered.
*$p < .05$, **$p < .01$, ***$p < .001$

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Table 8

Multilevel Models for High-Level Comprehension Outcome Variable Using Both Level-1 and Level-2 Predictors

	$Model_4$		$Model_5$		$Model_6$	
Variable	Est.	*SE*	Est.	*SE*	Est.	*SE*
Fixed Effects						
Intercept	3.04***	0.22	2.95**	0.32	3.01***	0.22
Reading Fluency	0.02**	<0.01	0.01**	<0.01	0.02***	<0.01
Grouping			0.17	0.45		
Piece 1	0.99**	0.31	1.56**	0.43	1.46***	0.37
Reading Fluency	<0.01	0.01				
Grouping			-1.13	0.60	-0.94*	0.43
Piece 2	0.90**	0.32	0.51	0.44	0.54	0.44
Reading Fluency	<0.01	0.01				
Grouping			0.80	0.63	0.89	0.61
Topic Knowledge	0.14	0.10				
Random Effects						
Intercept	0.75		0.88*		0.88*	
Piece 1	1.57*		1.43*		1.27*	
Piece 2	1.72*		1.85*		1.98*	
Within Student	2.15		2.07		2.03	
Deviance	765.14		743.92		743.50	
Number of estimated parameters	7		7		7	

Note. Text A and Text B variables not shown because they were not statistically significant level-1 predictors. Grouping was a dummy-coded variable with heterogeneous grouping coded as zero and homogeneous grouping coded as one.
Grouping was entered as uncentered. Reading Fluency was entered as grand-mean centered.
*$p < .05$, **$p < .01$, ***$p < .001$

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Table 9

Question Examples by Student Ability

High-Ability Students	Low-Ability Students
"What if Luke had stayed in the chemist business and hadn't named the clouds?"	"How long is Yosemite?"
"How would you feel if you got lost in Yosemite Park when you were there with your parents?"	"How many animals in the forest?"
"Why do you think the sequoia trees are called grizzly giants?"	"How high can sequoia trees go?"
"If you were Luke, would you work in the chemist shop or do what you love?"	"Why are people rowing on a raft in the river?"
"Have you ever read a 'Who Was' book like this where someone made a system that becomes popular?"	"Why are the trees gigantic?"
"Do you think a smokejumper's job is harder than a hot--a regular firefighter's job?"	"Are these types of clouds similar to the cloud types when we learned in our science textbook?"

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

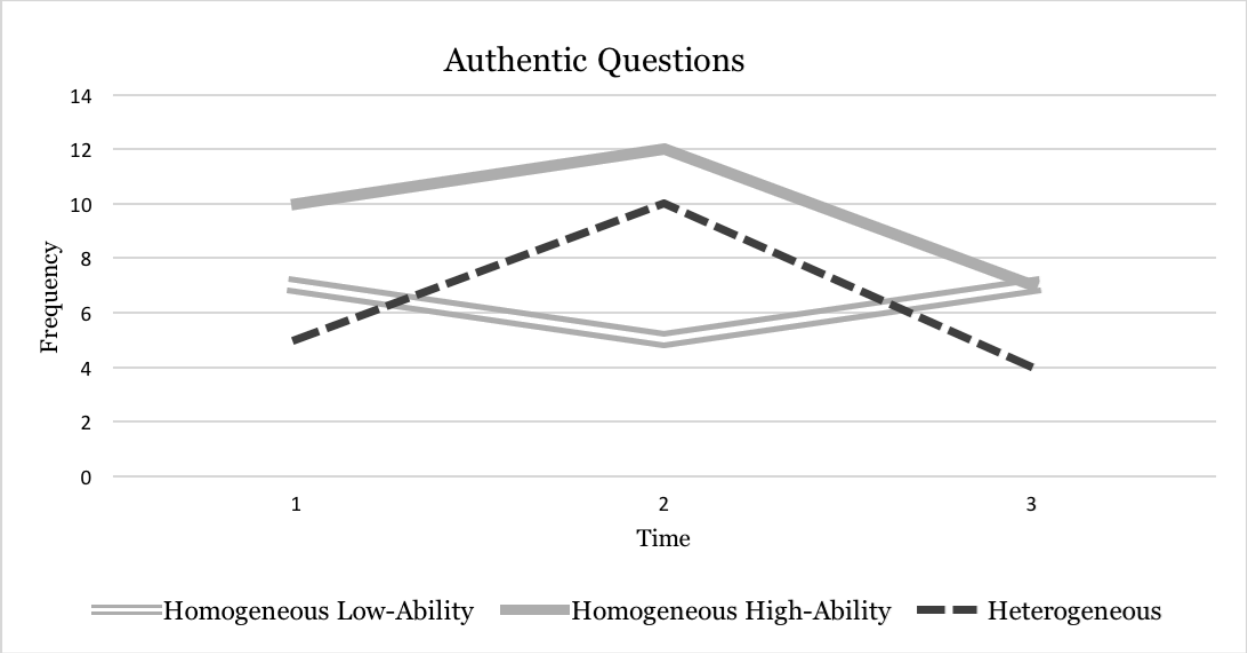

Figure 1. Depiction of the trends in the frequency of authentic questions over the duration of Quality Talk in fourth grade for the homogeneous low-ability group, the homogeneous high-ability group, and a heterogeneous ability group.

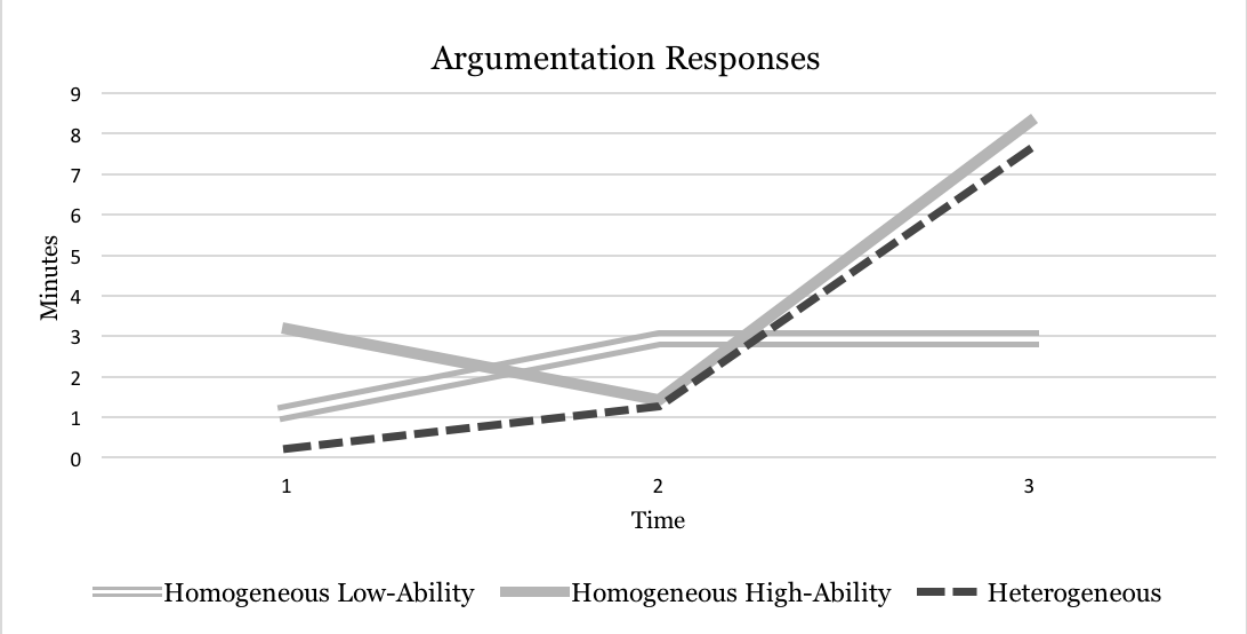

Figure 2. Depiction of the trends in the duration of argumentation responses over the duration of Quality Talk in fourth grade for the homogeneous low-ability group, the homogeneous high-ability group, and a heterogeneous ability group.

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

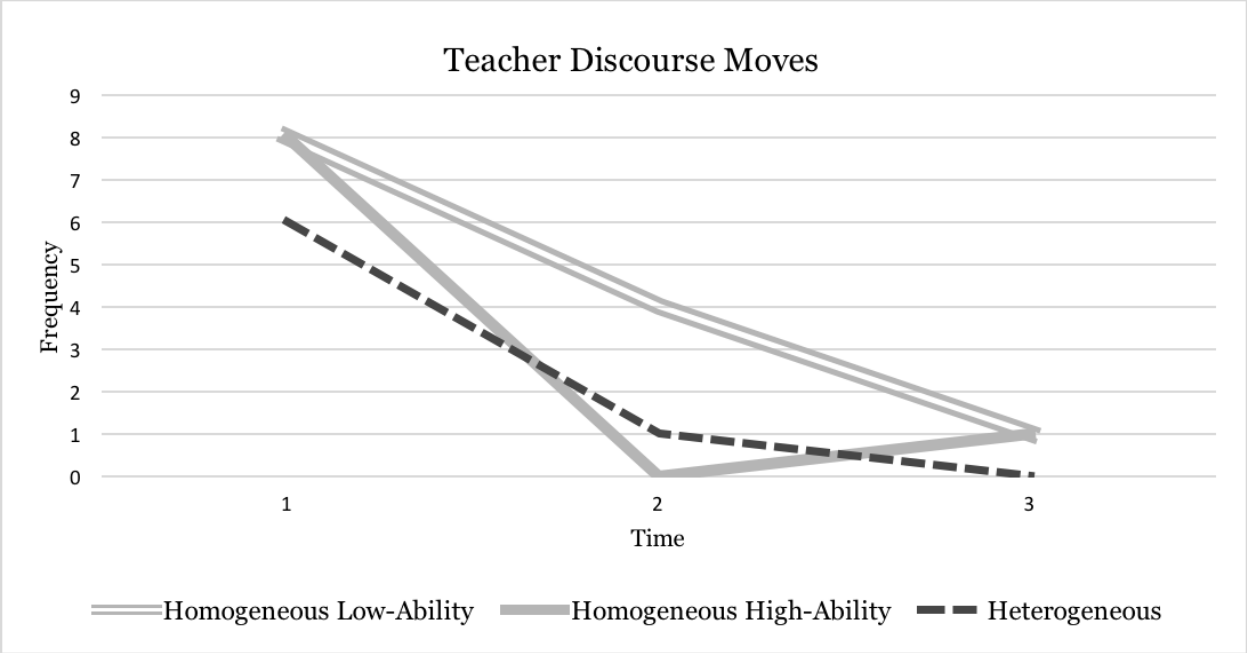

Figure 3. Depiction of the trends in the frequency of teacher discourse moves over the duration of Quality Talk in fourth grade for the homogeneous low-ability group, the homogeneous high-ability group, and a heterogeneous ability group.

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Appendix A

Discussion text: *The Dinosaurs of Waterhouse Hawkins*
5th grade, heterogeneous ability group, T2

[Start Time: 00:29]

Teacher: This is [Teacher]. Today's date is December 11th, 2014. The story we are discussing today is *The Dinosaurs of Waterhouse Hawkins*. The discussion group is [name]. Can you please go around and say your class and your number?

Student 1: Class 1, Number 1.

Student 2: Class 1, Number 2.

Student 3: Class 2, Number 3.

Student 4: Class 2, Number 4.

Student 5: Class 2, Number 5.

Student 6: Class 1, Number 6.

Teacher: Let me just check to see that we're all on camera, okay? There we go. Now we're all on camera. We have no students absent today. Our discussion group goal is going to be to focus on asking uptake questions that try to get us to have our peers give evidence to their claims and reasons. You guys are being really good at participating. Some are you are working on just listening and holding back a little bit on talking, and some of you are working on talking more. As a group, we're gonna work on asking uptake questions to try to find that evidence. "How do you know that?" "Where did you see that in the story?" "What happened to you before?" and "Prove it to me." Okay? That's what we're gonna work on in this discussion. The rules are that we don't need to raise our hands; we talk one at a time; we give others time to speak; we respect others' opinions; we consider or think about others' ideas; we give reasons to explain our ideas; we question and argue about ideas, not people; if we disagree we ask why.

[Time: 01:57]

Student 1: **Why did Waterhouse [inaudible]?** **S[AQ]**

Teacher: Say it one more time a little louder?

Student 1: **Why did Waterhouse build the models?**

Student 2: *Well, he wanted to show people prehistoric animals and things because nobody actually, well not many, people knew about it.* *EE*

Student 5: And-

Student 2: *Well, they did know about it but not exactly what they looked like.*

Student 5: Yeah.

Student 3: He wanted to get a picture in their head, like, this is what they might have looked like maybe millions of years ago.

Student 6: *He wa- I think its- he wanted to- I think he wanted to probably show scientists and help scientists with their work and- oh, in here it says, [reads] "He wanted to create such perfect models that anyone, a crowd of curious children, and leading scientists, and even Queen Isabella herself, could gaze at his dinosaurs to see the past."* *EE*

[Time: 02:59]

Student 5: **Do you think he was pretty close to what they actually looked like? Cuz when they show it to you, I mean, do you think he's pretty close?** **S[AQ]**

Student 6: Um-

Student 3: Maybe-

Student 6: *-I would say yes, because- Sorry, [Student 3]. Um-* *EE*

Student 3: It's, it's okay.

Student 6: *May- has anybody ever gone to a museum? Like, where there's dinosaurs a-*

Student 3: Yeah.

Student 6: *-and stuff? And we have a lot more research now because lots of people have found things, lots of people researched a lot more into it. And, he has some in here that when you look at, you kind of- whenever you look at it, it kind of even looks like one of the dinosaurs that would be at one of our museums now. So, I would say he probably was, just because if they still have them looking about like that in our museums, it would probably be pretty similar to what they were.*

Student 3: *I think he could've- I think maybe on, like, a couple of them, he might have got, like, on this picture [pointing to picture in book], it shows maybe this one [pointing], he got closer to think- seeing what it might have looked like than this one [pointing]. Cuz maybe this one was a different color, or something like that. I think some- certain ones he got really close to- getting to see the past in some other ones that wouldn't have looked like the ones that were millions of years ago.* *EE*

[Time: 04:30]

Teacher: I'm going to ask you guys an uptake question: **Do you think that it was important for him that they were accurate?** TM **T[AQ]**

Student 4: Yes, yeah.

Student 6: Um, to him it pro-

Teacher: [Student 4], do you want to go ahead and answer?

Student 4: *Um, yes, I do think it was important, cuz, if it wasn't accurate maybe they were, maybe the scientists would use, um, some of his stuff, like how he had the party, maybe some of the scientists would use his models for research, uh-* *EE*

Student 5: If it was accurate, you mean?

Student 4: Yes.

Student 6: Um, maybe-

Student 4: Like if it wasn't accurate.

Student 6: *-to explain to the scientists that it was accurate or just to, like, make it accurate. If you turn to page 398 and 399, he has bones everywhere that scientists have found and that researchers have found and they know that they're dinosaur bones. And if you can see, he's matching them up with different places on the dinosaur. So maybe, like, one or two spots weren't too accurate, but you can tell that a few of the spots are definitely pretty accurate because there- they even have pieces from real dinosaurs in them that will show a little bit more accuracy when you're trying to figure out.* *EE* ET

Student 4: Yeah.

Student 2: But how would scientists study that? Cuz it's not actually like real dinosaur bones cuz they- it was like-
Student 5: Yeah.
Student 2: -modeled of them, like clay models.
Student 5: Yeah.
Student 2: I know he found bones of them-
Student 5: Cuz you can't-
Student 2: But-
Student 6: He knew-
Student 2: But he- scientists can't really study off clay models. They have to, like, get bones and-
Student 5: Yeah, and there's, like, not really proof that that was completely accurate, and-
Student 2: Yeah.
Student 5: -even today we know that wasn't completely accurate, cuz-
Student 2: Mhm.
Student 5: -if they would have gone with that maybe they wouldn't of, like, known what they looked like- cuz today we know what they look like, but if they went with that, they could have just thought that-
Student 4: Mostly what they look like.
Student 6: *And don't you think that either way, if it was completely accurate or not exactly accurate, lots of people would still be impressed? Scientists would be impressed, even the Queen herself would be impressed, because nobody had done something this big, and for him to be able to try to get a picture in his mind and be able to make it, I think would be pretty impressive.*
Student 3: Yeah, I understand what you're- I understand what you mean.
Student 2: Me too.

EE

[Discussion Continued: 6:55 to 13:30]
Teacher: We're going to have to stop right there.
Students: Awww.
Teacher: Sorry, I know. I'm so sorry. Time is- we're going to have to watch out, with lunch and the time that we have. But I wanted to talk quickly about how we did. So, at the beginning our goal was to ask uptake questions and to provide evidence. How do you think you did with that?
Student 4: Pretty good.
Student 2: Yeah.
Student 6: Mmm.
Student 1: I think- yeah.
Student 3: [Nods]
Teacher: I think you did pretty well, too. I wrote some notes down, so I wouldn't forget. I had: [Student 6] used some evidence from the book. I had: [Student 1] use some stuff. [Students 2, 3, 4, & 5], everybody here either asked someone, "How do you know that?" or looked for evidence in the book. I am very proud of you for that. Why don't you think to yourself a goal for next time that you personally want to work on. You may go and do your writing now.
[End Time: 14:19]

Note: For readability and brevity, instances of excessively repetitive words or rephrasing (e.g., "um," "like," or "I think he wanted- I think he- I think he wanted to…") were either omitted or abbreviated (e.g., "I think he wanted to…") in the transcribed excerpt. Importantly, while omitted in the transcribed excerpt, all discourse was analyzed for the present study.

Discourse codes are indicated in the right column using the following abbreviations: **S: Student-initiated question**; **T: Teacher-initiated question**; **AQ: Authentic question**; *EE: Elaborated explanation*; ET: Exploratory talk; TM: Teacher discourse move. In the transcript, questions are emphasized in boldfaced type, individual argumentation responses (i.e., EEs) are emphasized in italicized type, co-constructed argumentation instances (i.e., ET) are emphasized by underlined text with the challenge component noted in a double underline, and teacher discourse moves are emphasized with wavy underlined text.

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., Wei, L., & Croninger, R. M. V. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology, 51,* 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003