Developing and trialing a measure of group thinking

Rupert Wegerif, Jonathan Doney, Taro Fujita, Richard Andrews, Julieta Perez Linares, Claire Van Rhyn.

Graduate School of Education, University of Exeter, Exeter, UK

Abstract:

This paper offers a critical review of the issue of assessing the quality of group thinking, describes the development of a Group Thinking Measure that fills a gap revealed by the literature and illustrates the use of this measure, in combination with interpretative discourse analysis, as a way of distinguishing those behaviors that add value to group thinking from those behaviors that detract value. The Group Thinking Measure combines two tests of equal difficulty, one for individual use and one for use by triads. This enables a measure not only of how well groups are thinking together but also a correlation between individual thinking and group thinking. This innovation gives an indication of whether or not working in a group adds value and so the extent to which a classroom culture supports collaborative thinking.

Key words: Classroom Dialogue, Group Thinking, Collective Cognition, Collaborative Learning, Teaching Thinking.

Introduction

Wegerif, Mercer and Dawes (1999) reported in *Learning and Instruction* on the use of Raven's non-verbal reasoning matrices to measure the success of an intervention on group thinking and also on individual thinking. The Standard Progressive Matrices Raven's test had been divided in half to make two equal tests, one made up of the odd questions and the other of the even questions. The two tests had been used before a three month intervention called Thinking Together which promoted Exploratory Talk and then again at the end of the intervention. One test had been given to groups using just one answer sheet and the other test, two days later, to individuals. A control study was used and the results indicated significant increases in both conditions. The argument of our paper was that an intervention designed to improve the way that small groups talk together not only had an impact on group ability to solve reasoning test problems but also on individual thinking as measured by this non-verbal reasoning test. We claimed that this finding supported the Vygotskian theory that individual thinking is mediated by internalized social interaction. Since its publication in 1999 that article has consistently featured in the annual most cited and most down-loaded lists of *Learning and Instruction* indicating that there might be some interest in the method we used and in the claims that we made on the basis of this method. This new paper returns to the principles of that method, locates it in more recent literature and develops an improved version of the same method in the form of a new Group Thinking Measure.

Asking groups to solve Raven's non-verbal reasoning tests before and after an intervention teaching Thinking Together was used as a way of assessing improvements in group thinking in a number of studies in the UK, Mexico, and, most recently, in China (Rojas-Drummond, Fernández, & Vélez, 2000; Fernández, Wegerif, Mercer, & Rojas-Drummond, 2001; Mercer, Wegerif & Dawes, 1999; Mercer, Dawes, Wegerif, & Sams, 2004; Yang, 2015). In addition, in a series of studies in South Africa, Raven's tests have been used to evaluate improvement in the thinking of individuals after teaching the Thinking Together programme (Webb & Treagust, 2006). The methodology of using two exactly equivalent Raven's reasoning tests, one for groups and one for individuals, was used in just three studies, two in the UK and one in Mexico (Wegerif, Perez, Rojas-Drummond, Mercer, & Velez, 2005: Wegerif, Mercer & Dawes, 1999: Mercer, Wegerif & Dawes, 1999; Wegerif, 1996: Rojas-Drummond, Pérez, Vélez, Gómez, & Mendoza, 2003). This approach to evaluating the effectiveness of an educational programme provided data on the improvement of individual thinking as well as the improvement of group thinking. Where the tests were commensurate and equivalent this methods offers a third measure which is the correlation of the scores of individuals, thinking alone, with the score of the group when they are thinking together.

It has been argued that there is no such thing as ‘group thinking’ - only the sharing of individual thinking. However, if some groups get better results than the highest individual in that group then this indicates that working together in a group may have added value to individual thinking. Some groups, of course, get worse results than the highest individual in the group indicating that the group social interaction may detract value.

In this paper we report on the design and use of a new Group Thinking Measure consisting of two short equally difficult tests of individual thinking and of group thinking to be given before and after an intervention in order not only to measure impact on individual thinking and on group thinking but also to explore further the potential of correlating individual scores with group scores. We report on our work in progress. The measure has been developed and trialed and further research projects have begun to explore potential uses in measuring the impact of interventions. We illustrate the possible value of this measure with the example of its use in two classroom case studies. The Group Thinking Measure enables us to identify groups which add value and to distinguish these from groups which detract value. This was used in combination with discourse analysis to provide quantitative support for qualitative claims.

The paper has three parts. The first part consists of a critical literature review into research on group thinking, especially group thinking in schools, bringing out limitations in commonly used methods and arguing that it might be possible to measure effective group thinking more directly. The second part briefly describes the principles we used to develop a measure of group thinking which offers the innovation of correlating individual and group thinking, as well as how we validated this test. The third part gives two illustrations of how the test can be used in a classroom to identify groups that add value and groups that detract value in order to support a more detailed qualitative analysis of group thinking processes. These examples are offered as work-in-progress of the development and initial trialing of a new method for assessing the impact of interventions which aim at improving group thinking.

Part 1: The assessment of quality in group thinking

There have been many studies of the quality of group work in classrooms and interventions designed to improve that quality. In an extensive review of these studies Howe and Abedin (2013) identify 67 evaluations of dialogue in classrooms. According to Howe and Abedin's analysis, the majority of studies since 2000 have been model-based which means that they assume a model of good dialogue and assess the impact of an intervention against this model. Testing to a model means specifying criteria of what counts as effective talk and then evaluating the extent to which the observed and recorded talk in a classroom changes as a result of an intervention in the direction of meeting these criteria. Examples of models of effective talk for group thinking include: 'Accountable Talk' (Michaels, O'Connor, & Resnick, 2008) 'Exploratory Talk' (Mercer & Littleton, 2007) 'Progressive Enquiry' (Muukkonen, Lakkala, & Hakkarainen, 2009), 'Quality Talk' (Davies & Meissel, 2015) and 'Collaborative Reasoning' (Resnick, & Schantz, 2015). While this approach can tell us how effective teaching has been in an intervention programme it does not, in itself, tell us if the dialogue itself is genuinely effective in supporting and enhancing group thinking. To show that the model is effective other research is needed. The 'testing to a model' approach to researching the quality of dialogue, when used alone, courts the danger of circularity or 'petitio principii' where what is claimed to be proven is already being assumed in the premises of the argument.

Models of effective dialogue such as Accountable Talk, Exploratory Talk, Progressive Enquiry and Quality Talk are justified in a variety of ways including critical literature reviews providing supportive arguments. Theories of how ways of interacting might work to build knowledge and understanding are central to each model. These theories differ greatly but each of these theoretical models of effective talk also claims support from empirical research. Two main kinds of research are referenced: inductive research that builds from observational studies of classroom dialogues and more deductive research which teaches a model of dialogue and tests this indirectly through its impact on curriculum area attainment measures such as scores on standard tests in Mathematics or Literacy. In the following three sections I will consider in turn, some of the theories behind models of effective classroom dialogue, indirect support through the testing of deductions from these models and support from more inductive research.

Theories used to ground models of effective group thinking

The range of theoretical traditions referred to in the literature on effective classroom talk is large and appears to be expanding rather than converging. Some research teams reference mainly Vygotsky (e.g Mercer & Littleton, 2007 and Michaels, O'Connor, & Resnick, 2008), some rely more on versions of Piaget (e.g Howe, 2013), some invoke Habermas's theory of communicative action (e.g Flecha, 2000), some apply theories of effective argumentation (e.g Forman et al, 1998; Kuhn, 2015), some look to Bakhtin (e.g Reznitskaya et al, 2009; Wegerif, 2011) while others unpack theories as to the nature of scientific progress (e.g Muukkonen, Lakkala & Hakkarainen, 2009). Referring to a range of different theoretical sources does not imply that these models of effective dialogue are necessarily incompatible but in fact there are clear fissures within this field of research which correspond, to some extent, to the different traditions of theory that are drawn upon. There is a long tradition of contrasting Vygotskian 'socio-cultural construction' approaches to Piagetian 'cognitive conflict' models with the understanding that these are, at the very least, different in their focus (eg Forman & Cazden, 1985). Matusov and Wegerif both contrast the practice of classroom dialogue based upon Bakhtinian theory to the practice that stems from Vygotskian theories (Matusov, 2011; Wegerif, 2008). Kutnick, Ota and Berdondini (2008) argue for a relational approach to understanding and teaching dialogue which contrasts to the focus on explicit verbal argumentation of others such as Kuhn (2015).

Synthetic papers claiming to provide an over-arching framework for understanding and researching effective classroom dialogue appear regularly (Mercer & Howe, 2012: Kuhn, 2015, Barron, 2003, Kutnick, Ota, & Berdondini, 2008). However, as Wegerif (2008) illustrates with a study of how ontological assumptions impact on this area of research, different theoretical frameworks can lead researchers to see quite different things in classroom dialogues. This makes the idea of a single synthetic framework for researching dialogue implausible. Before there could be a single shared framework for categorizing and relating approaches to effective classroom dialogue there would need to be convergence on a single understanding of dialogue at a more fundamental philosophical level. There is no sign in the literature of such a convergence happening. The number of theoretical perspectives applied to dialogue are multiplying with the addition of complexity theory (Osberg, 2009), new-materialism (Barad, 2009), rhyzomatic analysis (de Oliveira, & Machado, 2015) and more. The desirability of convergence on a single understanding of dialogue could be seen as contested by Bakhtinian understandings of dialogue that lay stress on the generative value for understanding of differences in perspective (Matuzov, 1996, White, 2015, Wegerif, 2013). Ironically perhaps, for this perspective, the most dialogic way to research classroom dialogues may not be to converge on a shared Bakhtinian dialogic understanding of dialogue but to maintain multiple diverse theoretical perspectives in play (Kincheloe, 2005).

Indirect empirical support for models of good group thinking

Models of effective classroom dialogue that lie behind pedagogical programmes (Accountable Talk, Exploratory Talk, Quality Talk, Progressive Enquiry, Philosophical Enquiry, Collaborative Reasoning and others) all, to our knowledge, claim to be supported by empirical research in addition to their claims to be based upon convincing theoretical considerations. Most commonly this refers to quantitative studies that demonstrate indirect impact on tests in curriculum areas. Mercer and colleagues claim that when Exploratory Talk was taught in Mathematics and Science over one year achievement increased (Mercer, Dawes, Sams and Wegerif, 2004). A recent study of the impact of teaching Philosophical Enquiry, found that it had a positive impact on mathematics and literacy scores (EEF, 2015). Davies & Meissel (2015) justify their model of Quality Talk through reference to a meta-analysis of 42 studies looking at individual student comprehension and learning outcomes in primary school settings. Similar evidence is offered for Accountable Talk (Michaels, O'Connor, & Resnick, 2008) and Nystrand's dialogic spells of interactive talk (Applebee, Langer, Nystrand, & Gamoran, 2003).

One issue with this indirect approach to measuring the quality of dialogues is that each study can only tell us about the kind of dialogue that worked in a specific context, for example the context of school mathematics, and cannot offer a general measure of the quality of the dialogue. Another issue is that while these studies help to establish the value of teaching for a particular kind of dialogue in a specific curriculum context they do not shed direct light on the causal processes which might explain why teaching for this type of dialogue proved successful. This means that they are unable to ground specific models of dialogue, as is so often claimed, since, if we do not know the causal processes, we do not know if other dialogue-based or non-dialogue based programs might not have had an equal or greater impact on learning in that specific curriculum area.

The controlled trial approach comparing the effects of a method of teaching dialogue to other methods or to 'traditional teaching' is of use for policy makers who have to choose between which approaches to invest resource in but is of limited value for research as it does not shed any direct light on the theoretical controversies between different accounts of causal processes linking dialogue and learning. There is a danger here of an element of circularity in that controlled trials can be used to claim that they establish the value of models of talk that are already assumed to be of value but without being able to question the assumptions behind those models directly. For example the randomized control trial of the impact of 'Philosophy for Children' (EEF 2015) appears to show an impact on Mathematics results but does not tell us if this was mostly due to the confidence inspired by the programme, the student questioning, the explicit reasoning, the relationship of trust promoted in groups, the 'philosophical' content matter of the enquiries or any variable or complex combination of variables found in the programme.

Inductive evidence for models of good group thinking

An alternative approach to empirical grounding is induction from observation of the learning and thinking occurring in small group dialogues. Barron (2009) uses both quantitative and qualitative description to characterize the features of successful as opposed to less successful groups after giving small groups a series of problems to solve. This inductive approach proved effective in finding, as Barron put it, those ‘patterns of interaction that are more productive than others for establishing a working joint problem-space that allows the group to capitalize on the resources available to solve problems and to learn from one another’. (Barron, 2009, p350). Barron’s approach involved close discourse analysis of the development of understanding in the talk of children combined with statistical comparison of coded features of successful and less successful groups. Close discourse analysis of the emergence of understanding in small groups is an approach that has been used very effectively by Mercer to ground some of the features of effective talk (Mercer, 2000). A similar inductive approach is also found in studies of online group collaboration leading to group cognition (Stahl, 2009).

Whilst a close interpretation of talk does often appear to be an effective way of revealing causal processes supporting the success of group thinking some implicit assumptions in this approach can be questioned. Methodologically this approach tends to focus on what is easy to see and record and analysis which is often the talk between students. Having assumed through the methodology that what is important is the talk, it is perhaps not very surprising if such studies often conclude that what is important is the talk. This may well be the case of course and such studies can distinguish between effective talk and ineffective talk. However the problem is that if there were key causal processes that were anything other than talk, invisible changes in neural pathways for example, or unspoken emotions then these cannot easily be found out by this method.

The kind of analysis of cognition in talk that Mercer refers to as socio-cultural discourse analysis (Mercer, 2007) depends upon a theoretical perspective often referred to as neo-Vygotskian which assumes that cognition is mediated by cultural tools especially language (Wertsch, 1998). The argument can at times seem suspiciously circular: first it is assumed that group thinking is found in patterns of language use, then patterns of language use are observed that correlate with successful problem-solving next it is claimed that those patterns ‘caused’ the success that they are associated with and finally it is claimed that these patterns of language use should therefore be taught. This critical challenge is a version of an argument first put forward in the 18th Century by Hume (1965/1751) who argued that induction is not a viable means of shared knowledge construction since we only observe correlations and not causes. Causation needs to be projected into the data by the imagination and the danger is that what we project into the data depends upon the theoretical assumptions that we start with rather than anything actually emerging from the data. This probably explains why close discourse analysis of exactly the same data by researchers from a range of different theoretical traditions can lead to a range of different interpretations of the causal processes behind the thinking found in that data (Koschman, 2011). The problem is that each research tradition will focus on different aspects of the data and see different causal processes. This skeptical challenge to induction as an approach does not invalidate all the research done in this tradition, any more than Hume's skeptical challenge in the 18th Century invalidated all the claims of empirical science to find causal laws from observations. It just suggests a certain caution in the claims we make and the need for awareness of the extent to which findings might derive from the assumptions that researchers bring to the data as much as from what they find in the data.

Inspiration from outside the field

So far we have argued that the methodologies applied in research on classroom talk can be challenged. There is some inductive evidence of processes that appear to be effective in group work and there is some indirect evidence of effective pedagogy promoting group thinking gained from measures of success on various tasks, usually curriculum related. Recently a claim has been made from outside the field of educational research that group thinking can be measured more directly. Woolley and colleagues have pioneered an apparently inductive statistical approach to assessing group thinking (Woolley et al, 2010). A range of different groups ranging from two to five members, were given a variety of different kinds of tasks drawn from the McGrath Task Circumplex, which Woolley et al describe as: 'an established and validated taxonomy characterizing tasks according to the dominant coordination process required for its accomplishment by a group'. These tasks included generating tasks, choosing tasks, negotiating tasks and executing tasks. Factor analysis of the results strongly suggested that some groups thought together better than others across all these different kinds of task leading to a construct for collective intelligence referred to as 'c'. 'c' could predict the group's performance on new tasks better than any measures of the abilities of the individuals making up the groups including average cognitive ability of group members or measures of the highest cognitive ability within the group.

Interestingly one of the tasks used was an almost exact copy (unattributed) of the evaluation method used by Wegerif, Mercer and Dawes (1999). A standard Raven's progressive matrices test was divided into two equal tests by taking all the odd questions for one test and all the even for the other. One test was then given to groups to solve working together and the other, separately, to the individuals who made up the groups. This was one of the ways that the team demonstrated that the group cognitive ability did not correlate closely with the cognitive ability of the individuals making up the group. Of all the tasks used to test the groups the one that correlated most closely to 'c' was the Raven's test at 0.86. This means that, while group intelligence or 'c', is a product of measuring performance on a range of tasks if one were to choose only one task to measure it, that task should be something like the graphical puzzles used by the supposedly culture-free because 'non-verbal' Raven's reasoning test.

This inductive but quantitative approach to researching group process did not assume one model of good group work yet was able to say something interesting about the characteristics of more successful groups. Several factors that might have been thought to correlate to 'c' were tried and discarded including measures of motivation, group cohesion and satisfaction. Only three factors emerged as significant, the presence of women, 'social sensitivity' measured using the 'reading the mind in the eyes' (RME) test, and also the distribution of turns at talk where groups with a more equal distribution did better (Woolley, 2010). The significant impact of the presence of women in groups was largely explained by the fact that women showed greater social sensitivity leaving just two key factors for successful group thinking: social sensitivity and equal turn distribution. Although the test used to measure social sensitivity seems to assume a face to face situation as it involves interpreting pictures of facial expressions, further studies have demonstrated that this same test predicts group success even in online-mediated group tasks (Engel et al, 2015). This implies that the test succeeded in accessing the ability to read intentions and put themselves in the position of others (so called 'theory of mind') even when not face to face.

A possible criticism of the Woolley et al approach to measuring group thinking is that it is in fact indirect since the measure is of success on various tasks rather than a measure of the quality of the thinking in itself. The tasks measured by Woolley et al. were relevant in an enterprise context but did not include all the kinds of group thinking relevant in educational contexts. It is conceivable that some kinds of group thinking considered to be of high quality, reflecting on assumptions in a philosophy seminar for example, could be experienced or recorded without leading to improved performance on any of Woolley et al.'s tasks. In response to this criticism we propose that what Woolley et al. refer to as the measurement of the quality of group 'intelligence' be referred to as a measure of 'effective group thinking' meaning the kind of group thinking that can be effective in a range of group tasks.

Discussion of the assessment of the quality of group thinking

The 'testing to a model' approach already assumes a model of effective group thinking and so can only assess whether or not the pedagogy achieved the desired result without being able to assess the value of the model itself. Indirect tests of group thinking through the results of curriculum assessments can tell us if pedagogical approaches to teaching dialogue are effective but, in themselves, they cannot tell us about the causal mechanisms that are general to effective group processes. Inductive qualitative approaches that draw conclusions directly from the observed data of groups talking together and solving problems together can appear to reveal causal processes that link the way students talk together to learning outcomes. However, these approaches can be criticized for potentially smuggling in assumptions about good group thinking processes that shape what is observed and what is not observed. The psychometric approach to measuring group thinking adopted by Woolley et al can also be criticized for its assumptions as to what constitutes good group thinking but it offers a more direct measure of effective group thinking than any curriculum assessments. This makes it a potentially useful addition to the range of approaches used to assess the quality of effective group thinking.

The study by Woolley et al has supported the relevance of using Raven's type non-verbal reasoning test problems to evaluate group thinking. However, the inductive statistical approach used by Woolley et al is too distant from the actual interactions that carry the causal mechanisms of group effectiveness to say much about these beyond the importance of social sensitivity and balanced turn-taking. Inductive qualitative approaches like those of Barron and Mercer seem to be able to say much more about causation but run the risk of assuming too much about what good group process is. The combination of a direct quantitative measure of effective groups with qualitative analysis of the processes underlying that quantitative result could be the best way to mitigate the various weaknesses of the different approaches. The implication of this critical literature review is that a relatively 'objective' quantitative approach to measuring group thinking with standardized tests should be used in combination with interpretative inductive methods to measure the quality of group thinking.

Part 2: The development of a Group Thinking Measure.

Evaluations of the impact of teaching 'Exploratory Talk' in the UK, Mexico and China have used Raven's non-verbal reasoning test problems to assess the thinking of groups in designs which also measure the thinking of individual students (Wegerif, Mercer & Dawes, 1999; Mercer, Wegerif & Dawes, 1999; Wegerif et al, 2005; Mercer & Littleton 2007, Yang, 2015). The virtue of the methodology used is that it is possible to video groups working together around specific visual puzzles that they manage to solve or not. This leads to qualitative insights into features of successful group process (Wegerif, 2007 chapter 4 gives several exemplifications). This method has been used within pre and post designs to assess the impact of educational programs on the talk of the children. Although similar visual puzzles taken from Raven's series of tests have been given both to groups and to individuals there have been problems with the effectiveness of the use standard non-verbal reasoning tests for measuring group thinking. When children are asked to talk together full tests take longer and also tend to lead to ceiling effects on the Standard Progressive Matrices. It has not yet been possible to cross-correlate in the studies between the scores of individuals and the scores of groups to see which groups are providing extra value, which remain the same as the highest individual in the group and which are detracting value through producing a lower score than that of the highest individual.

To remedy these weaknesses a new method has been developed called the Group Thinking Measure. Two short sets of visual puzzles have been created to be used in combination. The overall measure of Group Thinking combines a measure of individual thinking correlated to a measure of group thinking with a measure of the difference between the individual scores and the group score. The Group Thinking Measure as a whole is intended to provide insights in whether or not working as a group is adding value by getting a better result than any individual alone, or detracting value. This is useful because it can assess whether the culture of the classroom promotes collaborative work or not and so can support teachers in changing the culture of their classroom. Like the previous use of Raven's test problems this new Group Thinking Measure is particularly useful in integrating qualitative interpretations of group processes using videos of groups working together around the tests, with quantitative measures of the success or failure of group thinking.

Thirty test problems, making up two similar tests of 15 questions each, were designed by Dr Andrew Richards and drawn by educational researcher and graphical artist Claire van Rhyn. Following the successful model of Raven's matrices each puzzle has a grid of nine shapes with one missing (see Figure 1). The participant has to select the correct shape to complete the pattern from eight options. No language is required to understand the test problems. The tests range from early problems that require only simple pattern recognition and completion to latter problems with combinations of several different manipulations.

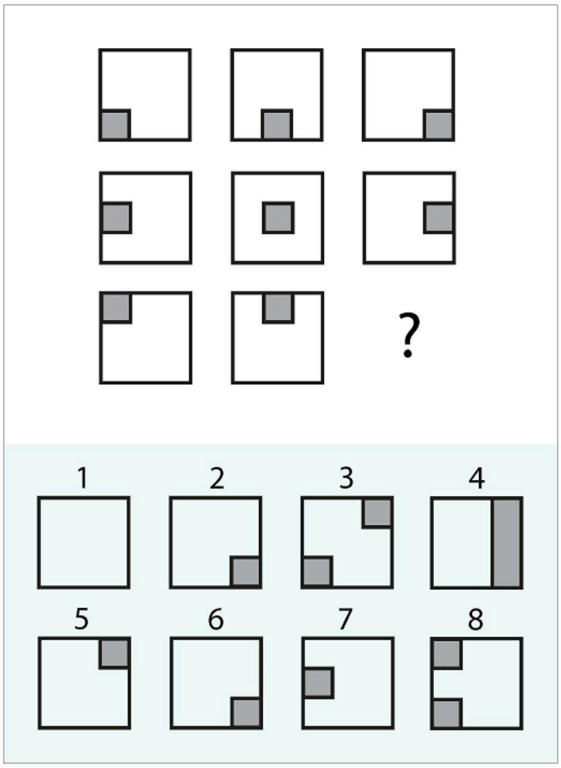


Figure 1 One problem in the Group Thinking Measure

The Exeter network of accredited ‘Thinking Schools’ who are interested in such assessment tools were asked to help validate the test. 220 pupils aged 9 to 11 from 6 different schools in the UK and South Africa participated. The cultural and socio-economic range was diverse. Half the classes used test A as individuals and test B as groups while the other half did this the other way around. Half the classes did the individual condition first and then the group and the half the other way around. To test for the similarity of the two tests 102 pupils were split in half, one half doing test A and then test B while the other half did test B and then test A.

It was found that the order in which the tests were done had no significant impact on the results (ANOVA , F = 0.367, p=0.546). A paired samples t-test was used to test the hypothesis that there would be no significant difference in the test score between the two tests if they were of equal difficulty; this showed that there was a statistically significant difference in the scores for test A (M=8.75, SD=2.58) and test B (M=7.30, SD=2.41); t (43) =4.182, p < 0.001. Thus it was concluded that the original versions of test A and B were not of equal difficulty.

As a result, the puzzle tests were reallocated to ensure equality between A and B, and to ensure that the puzzles became progressively more challenging.

The redesigned scales were tested in the same way as the original tests; a group of 30 mixed ability year 7 students were randomly allocated to one of two groups, each comprising 15 students. One group undertook test A, and the other, test B.

Once again, an analysis of variance between groups indicated that the order in which the tests were done had no significant impact on the results (F = 0.096, *p* = 0.757).

The percentage of correct responses for the revised tests A and B is shown in table X below.

Table X: **Percentage of correct responses for revised tests A and B**[1]

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	98	95	86	86	79	74	56	56	49	35	33	26	23	12	9
B	98	93	91	84	79	74	67	51	35	35	33	26	21	12	12

Again, a paired samples t-test was used to test the hypothesis that there would be no significant difference in the test score between the two tests if they were of equal difficulty; this showed that the difference in the total scores for test A (M=8.16, SD=2.22) and test B (M=8.09, SD=2.45) was not statistically significant (t (42) = 0.231, p = 0.819). This suggests that the revised test A and B are of equal difficulty.

Reliability analysis was undertaken for the revised tests, each of which consisted of 15 items; Cronbach's α for the revised tests were: test A α = 0.54, test B α = 0.65. Thus the reliability of the tests appears to fall below the often cited threshold of 0.7 (Nunally 1978). However, the dominance of Cronbach's α has been under increased scrutiny (e.g. Sijtsma 2009). Guttman's λ_6 (smc) was also calculated for the revised tests, showing A λ_6 = 0.69, test B λ_6 = 0.77. Based on these data, the tests were considered to be of sufficient reliability to take forward to the next stage of development. Further analysis will be undertaken on the full data set in due course.

Perhaps the main innovation of the Group Thinking Measure is that it makes it easy to correlate the individual scores with the group scores. It is clear that if all the individuals in a group score low scores and then, when working as a group on a test of equal

[1] The two full tests are available on: https://socialsciences.exeter.ac.uk/education/research/centres/teachingthinkingdialogue/cedu/cognitiveprogrammesandtools/groupthinkingmeasure/

difficulty, they score a much higher score, this is likely to be due to the value added by working as a group. The same effect can happen the other way with individuals scoring high score working alone and a low score when working as a group. In theory a good group should be able to share ideas and so should score at least as highly as the best individual in the group would score. We therefore decided to take a group score of at least one standard deviation higher or lower than the highest individual score as a measure of significance. Groups that score over a standard deviation more than the highest score of any of the indivduals in the group are defined as Value Added Groups, groups where the group result is more than one standard deviation lower than the highest individual within the group are called Value Detracting Groups and groups that score between these two are defined as Value Neutral Groups. Because we are dealing with each group as a separate case we know the full population and use the population standard deviation calculation and not the sample standard deviation calculation.

We trialled the revised Group Thinking Measure with 29 children aged 11 and 12 in St. Mary's Catholic Primary School, Bridgend, South Wales (accredited as an advanced thinking school by Exeter University). Using the standard procedures of research in 'Thinking Together' (e.g Wegerif, Mercer & Dawes, 1999) the children were arranged by the teacher in 8 mixed gender and mixed ability groups of three with one mixed gender mixed ability group of 4. The mixed ability was determined by the teacher using her knowledge of the children's academic ability. The results suggested that the measure was effective in distinguishing between the groups finding 3 Value Adding Groups, 4 Value Neutral Groups, and 2 Value Detracting Groups in the classroom. This information was of interest to the teacher in supporting her efforts to promote a collaborative reasoning culture in the classroom.

Part 3: An illustration of using the Group Thinking Measure to frame qualitative studies.

The main projected use of the new Group Thinking Measure is for evaluating the impact of educational programs, such as 'Thinking Together', which aim to teach effective dialogue. This is similar to the use of a modified version of Raven's Standard Progressive Matrices made earlier by Wegerif, Mercer and Dawes (1999, see also Mercer, Wegerif & Dawes, 1999). One value of combining two equal tests is assessing not only the thinking of groups but also the impact that teaching dialogue has on the thinking of individuals. In this way it is possible to observe if the thinking of individuals is improved by teaching better thinking in groups. Following the findings reported by Wegerif, Mercer and Dawes (1999) we hypothesize that individual non-verbal reasoning is dialogic and therefore is likely to improve as a result of improvements in group thinking. When both group thinking scores and individual thinking scores increase as a result of teaching dialogue the difference between the mean of the individual scores and the group scores might not change. This means that positive increase in the difference between the mean score of individuals and the scores of the groups cannot be, used alone, a direct measure of effective group thinking. Nonetheless the difference between individual scores and group scores can be a very useful indicator of the culture of classrooms. A large number of Value Detracting Groups in a classroom indicates that action should be taken to improve the culture of collaboration. A large number of Value Adding Groups might indicate that action needs to be taken to help student transfer dialogic thinking strategies from their group work to their individual work.

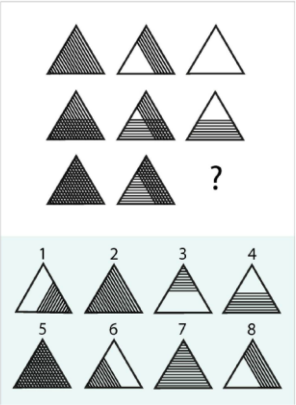
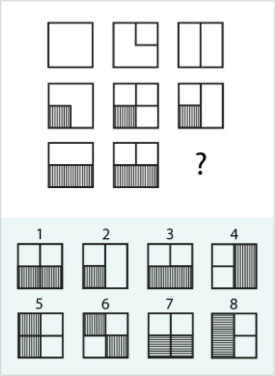
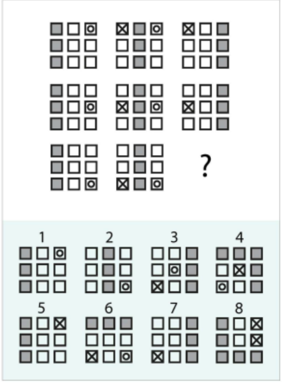
The most useful way to use the Group Thinking Measure is in combination with the analysis of videos of students talking together to solve the group thinking test. Exploring the ways in which children fail to solve problems as well as the ways in which they succeed in solving problems can be of formative value. With this approach, using the Group Thinking Measure as a pre and post assessment of an intervention teaching dialogue, it is possible to see how the same group of children can learn to solve complex puzzles by working together that they previously failed to solve. As with the earlier use of modified Raven's reasoning tests (Wegerif, Mercer & Dawes, 1999) the value here lies in combining a quantitative measure of success and failure with qualitative analysis of the reasons for that success or failure. This clear evidence of the impact of interaction strategies is motivating and supportive for students and teachers.

To illustrate the potential of the Group Thinking Measure in combination with qualitative video analysis to isolate value-adding behaviors and value-detracting behaviors we conducted two cases in classrooms in schools in the South-West of the UK. In each case we used the test results to focus in on successful groups solving problems and contrast these to less successful groups failing to solve problems. Without the test results we could not have been sure that the strategies and behaviors we thought were good were actually effective in solving problems.

Case study 1: Characteristics of successful group work

As a pilot study, we worked in a class of 35 pupils aged 10 and 11 (18 girls and 17 boys). The teacher put this class together into 11 groups of three and one pair. Each group was mixed gender and mixed ability decided by the teacher based on her knowledge of their academic performance. This followed the standard procedure in all the previous work on 'Thinking Together' (e.g Mercer, Wegerif & Dawes, 1999). We selected 6 of the groups at random to video working together around the original test B. In this earlier version of the Group Thinking Measure test B turned out to be significantly harder than test A so we were unable to do the full analysis of whether groups added value to the individual results or not. However the test results did enable us to distinguish the more successful groups from the less successful groups. To contrast less successful with more successful group work we compared Group 6, who scored 7 out of 15, with group 7, who scored 12 out of 15. For each group, we began by dividing the problems into those that were solved correctly and those that were not solved correctly. These tests were done on paper as the class did not have access to enough computers to do them online. The analysis was done directly from the video writing notes and codes on paper without transcription. The results below are presented using the simple 'insightful observation' method pioneered by Barnes and Todd (1977). Like all interpretative research this approach assumes an element of participant observation in which the researcher imagines themselves vicariously as a member of the group and explicates intuitions as to what is going on (Habermas, 1977).

Problem	Group 6 (7/15)	Group 7 (12/15)

B6 	Incorrect One pupil said there is a pattern but then did not say what it is. They pointed to answer 4 or 7. One pupil said "I think 4, because, lines" and all agreed without any exploration. There were no shared smiles or laughter.	Correct The group began by systematically eliminating all the impossible answers and then one of them found the answer (7), another pupil said 'I don't understand' and the other two explained why. Their fingers were always on the paper together moving to point out different alternative (see Figure 1). One of them smiled after they agreed with the answer.
B9 	Incorrect 2 was suggested by one pupil but this was rejected because it was the same as one of the patterns. Then 5 was suggested by the same pupil who rejected 2. The third pupil agreed giving the argument that the shaded parts were extended (incorrect analysis). They agreed without further explorations. There were no shared smiles or laughter.	Correct When they looked at the problem they smiled ironically together because they thought that it looked hard. They confirmed "None of them are the same, they are all different", with shared laughter. They then compared 1 and 7, and then decided on 1 as it has vertical lines (which is correct). Their fingers were always on the paper. After solving the problem, one of them noticed that not everybody had a pen this made them all laugh again.
B10 	Incorrect All pupils just throw their thoughts. They noticed the correct positions for a circle and a cross but did not comment on the more salient shaded pattern. They chose 6 based on this. There were no shared smiles or laughter.	Correct The group began by systematically eliminating all the impossible answers. They noticed the shaded patterns and chose 4 and 7. Then they changed their mind and chose 3, 'because x has to be a corner' as one of them said. Then another pupil suggested 7, but the third pupil disagreed. She reasoned 'look, not one of them (in the third column) has a circle'

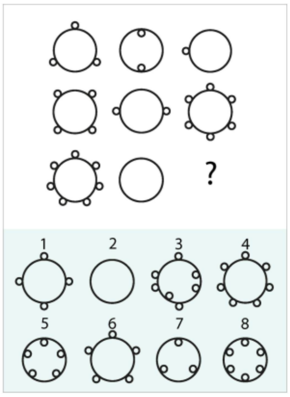
		and then they all agreed 7. Their fingers were always on the paper, and they smiled at each other several times during their problem solving. They again smiled together when they had solved the problem.
B13  	Incorrect They started counting numbers of small circles. One pupil noticed circles can be outside or inside bigger circles, but the other two did not comment. After some silence, 4 was suggested, but one child ignored this and then started counting the numbers of small circles again, and then found 1, 2, 3, 4, 6, and 7, but 5 was missing. One child pointed out there were 'two fives, inside (5) or outside (6)' but the other one just said "Yes, no. 6" without any confirmations, and then the third one agreed. Only one pupil put her/his fingers on the paper. There were no shared smiles or laughter.	Correct They could not solve this question at their first attempt. They asked the teacher whether they could come back to this one later. After they had finished all the other questions they came back to this problem. They then started noticing additional elements in the problem. One pupil suggested addition, but then the other pupil suggested outside is addition but inside is subtraction, and then noticed 7+0=7, and decided that the answer is 4, which is correct. Their fingers were always on the paper. They finished with big smiles apparently pleased to have solved this together.

Table 1: comparing successful and unsuccessful group work

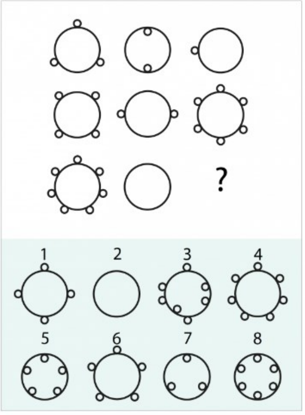
Figure 1: In solving the printed version of the text the presence of all hands on the paper seemed significant

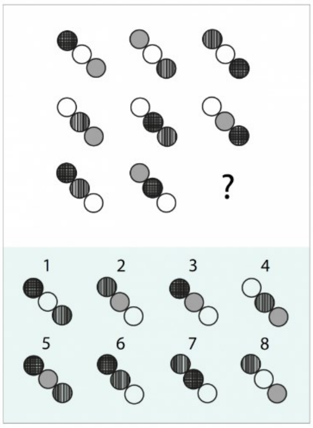
Case study 2: Characteristics of Value Adding Groups

In the next case study with 15 students aged 10 and 11 we were able to use the revised Group Thinking Measure and so we were able to calculate the difference between the individual and the group score.

Group	Individual scores	Group scores	SD (Pop)	Type
1	10,8,6	9	1.632993	VNG
2	8,5,3	11	2.054805	VAG
3	14,7,4	9	4.189935	VDG
4	8,5,5	9	1.414214	VNG
5	7,7,3	5	1.885618	VDG

Our analysis found one Value Adding Group and two Value Detracting Group. Using the same type of analysis as the previous case study we compared the questions that the Value Adding Group had succeeding in solving that the Value Detracting Group had failed to solve.

Problem	Group 2 VAG	Group 3 VDG
B11 	Correct Started with a shared smile. One of them immediately suggested 4, which was a correct answer. The other agreed.	Incorrect There was silence at the beginning and two of them put their hands on their mouths. One of them suggested '1, 5, or 2'. No responses from the other two. The same pupil prompted them again with, 2 or 5 and they chose 5 without further reasoning.

B13 	Correct One of them suggested 2, and the other one 7 or 2, stating this is a guess (with laugh). Then another said “May be… 7, because two lines…”, but did not reason further. Then one of them suggested that ‘No, actually this one is 2”, pointing the patterns which he noticed with fingers. This suggestion was agreed and they finished this problem with smiles.	Incorrect 4 was suggested by one pupil, and then this was agreed without further explorations.

Why analyzing talk alone is not enough

We were initially surprised when our Group Thinking Measure showed group 2 to be the most Value Added Group. This is the group that talked the least completing the questions faster than any other group. However, when we compared the group dynamics between this group and the most Value Detracting Group, group 3, there was a marked difference. This was most evident not in the talk but in the body language. Figure three shows poses typical of the two groups. While group two tended to smile together, laugh together and orient their bodies towards each other group 3 tended not to be very responsive or interactive with no laughter and few smiles, hands over mouths and bodies kept apart in separate spaces.

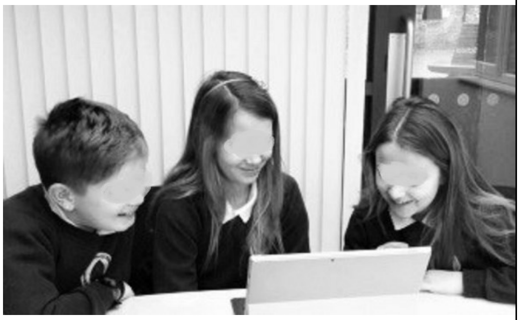
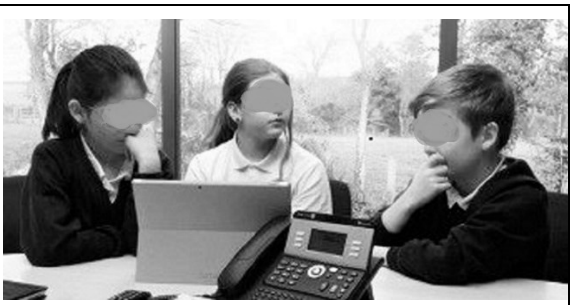
	
VAG	VDG

Figure 3: Comparing body language between a VAG and a VDG.

If we look again at how a key question was solved, B13 illustrated above, this was solved by the one boy in the group who changed his mind saying ‘no … actually it is 2’. As he did so his previous frown of concentration turned to a big smile. It was clear that his understanding had changed. This change was not directly co-constructed through talk with the other two members of his group. But this does not mean that the talk and the group work was irrelevant to his ability to solve the problem.

Solving these kind of puzzles involves spotting a pattern that is not immediately obvious. As with the boy in group two, this can involve an ‘Aha!’ experience of the type studied by Kounios and Beeman (2009) using brain imaging. Before the insight surfaced into consciousness Kounios and Beeman were able to observe it preparing in the brain so that they could tell when someone had an insight solution to a puzzle before the person themselves. As the ‘Aha moment’ occurred they found that people tended to blink more than normal, even closing their eyes and stilling their bodies. They hypothesized that it was necessary to reduce external stimulation for a moment in order to be able to listen to the relatively quiet internal voice bring them the solution. This is something that we have observed also in studies of problem solving in Mathematics and described as a dialogic switch because it is a switch in perspective in relation to an inner or invisible voice (Kazak, Wegerif & Fujita, 2015). In the video this switch in perspective is apparent with a sense of tension in the face of the boy being transformed into excitement. He suddenly says: ‘Actually no … I think it is 2, because look’ [he leans over and points the shapes on the screen] ‘if that one’s that then that one’s that,’ He continues. The two girls in the group both smile with a certain release of tension and say together ‘Oh yeahhh’ indicating that they understood his explanation.

Kounios and Beeman (2015) have found that relaxation and laughter facilitate insight creativity and that anxiety inhibits this. The solutions to the problems in group 2 were not co-constructed together using explicit talk but the positive group atmosphere with shared smiles and laughter might have created the sort of shared ‘dialogic space’ (Wegerif, 2013) in which it was easier for creative solutions to the puzzles to surface.

Thematic analysis of successful groups

Thematic analysis was conducted on the video data of all the videoed groups in both case studies, 11 groups in total. We focused on behaviors that appeared in the successful and Value Added Groups when solving problems but that did not appear in the unsuccessful or Value Detracting Groups when failing to solve problems. The analysis consisted of watching and coding the videos where any aspect of behavior that seemed potentially related to successful problem solving. Inevitably this coding was influenced by our knowledge of the literature. After an initial coding we iteratively returned to merge similar codes, delete outliers and delete codes of behavior that occurred equally in unsuccessful groups failing to solve problems as it did in videos of successful groups solving problems. This analysis was only made possible by the use of the Group Thinking Measure which enabled us to correlate our observations with a quantitative measure of group success and failure. This procedure revealed general behaviors that characterized successful problem solving:

- Encouraging each other, for example responding to suggestions with 'could be …'
- Expressions of humility, for example 'I do not understand this.'
- Giving clear elaborated explanations, for example 'the triangle here is removed and here it turns around by 90 degrees'
- Equal participation with everyone in the group actively involved in each problem.
- Actively seeking agreement from others, for example by asking 'do you agree?'
- Not moving on until it is clear that all in the group understand for example asking 'I do not understand it, can you explain again?'
- Open questions, for example 'can anyone see a pattern here?' and 'what do you think?'
- Warm positive affect with shared smiles and laughter.
- Willingness to express intuitions, for example, 'I am not sure but I have a feeling it is that one'
- Indications of mutual respect in tone and responses.

- Taking time over solving problems seen in accepting pauses and giving elaborated explanations when asked.

It is true that many of these features are found in models of successful talk used already as a basis for classroom pedagogy such as 'Accountable Talk' (Michaels et al 2008) or 'Exploratory Talk' (Mercer, 2000). The use of the GTM has enabled a more inductive approach in which features of successful groups can be derived from observation and analysis of how VAG solve problems in contrast to observations and analysis of how VDG fail to solve problems. Some of the features that have emerged from this initial case study such as the value of humour and the importance of expressing intuitions even when there is no supporting reasoning, are not found in all existing models of good group talk. Closely linking qualitative analysis of the interaction in groups to a quantitative measure of success in group thinking forced us to question the assumption that the group thinking is visible in the talk and to look for further explanations of group thinking in invisible neural processes stimulated and supported by group relationships.

Discussion and conclusions

A critical review of the literature relating to the measurement of group thinking raised several concerns. Most research on classroom dialogue assumes a model of good dialogue and then looks for features that correspond with that model. This approach might be useful to assess the effectiveness of the teaching of dialogue but does not measure the effectiveness of the group thinking. Research that claims to measure the effectiveness of teaching group thinking through the impact that this has on curriculum tasks is obviously valuable but it is also obviously indirect as a measure the effectiveness of group-thinking in itself. This kind of research can demonstrate the value of a pedagogical approach but does not shed light on the causal processes that might underlie effective group thinking. Inductive approaches such as interpretative discourse analysis hold out more promise of describing causal mechanisms in talk. Work by Barron and Mercer, as well as others, has led to useful knowledge about effective group processes and how to promote these. However, such approaches run the risk of being influenced by theoretical and methodological assumptions.

An inspiration for how we mind construct a more direct measure of effective group thinking can be found in the recent work of Woolley and colleagues which suggests the existence of what they referred to as a group intelligence factor named ‘c’, making some groups more effective than others across a range of problem-solving tasks. The single group task that correlated most closely to this ‘c’ was found to be a Ravens reasoning test. Our argument is that this more quantitative approach to directly measuring the effectiveness of group thinking could, when used in combination with qualitative research, mitigate the danger of distorting assumptions and help us get closer to uncovering the actual causal processes that lie behind effective group thinking.

Woolley et al’s approach partly reproduced an approach to assessing group thinking already pioneered by Mercer and Wegerif in several studies (e.g Wegerif Mercer & Dawes, 1999). This is to divide a standard Raven’s reasoning test into two equal halves and give one to groups and one to individuals. Woolley et al used this to show that group thinking did not closely correlate with the ability of individuals within those groups. They found that group effective thinking correlated more closely with social sensitivity but their quantitative methods did not allow them to drill down further to explore the causal processes that lay behind this finding.

In response to the limitations of qualitative inductive approaches to exploring group thinking on the one hand and the limitations of quantitative inductive approaches on the other we developed a Group Thinking Measure that can combine the two approaches. This new measure is a development and refinement of the approach already used by Mercer and Wegerif in the past and taken up and revived by Woolley et al, albeit unconsciously. Unlike the approach of Woolley et al this test can be easily used by teachers in classrooms as an aide to assessing the quality of group thinking and as a support programs teaching effective dialogue. The close correlation of scores on Ravens’ reasoning tests to ‘c’ found by Woolley and colleagues (0.86) suggests that our test, which is similar to Ravens’, might be a useful proxy for a more comprehensive test.

In addition to being an instrument that can measure the impact of teaching on group thinking and on individual thinking it also holds out the prospect of being a useful indicator of which groups are adding value to shared thinking and which groups are detracting value. We illustrated that use through a case study which applied the Group Thinking Measure to focus in on group behaviors that added value, contrasting these to behaviors that detracted value. While these confirmed many of the findings of earlier studies such as Mercer, Wegerif & Dawes (1999) the use of the quantitative measure also pushed us to account for thinking processes that were not visible in the talk of children but that could be inferred from their behavior. Exploring these processes and also processes of communication visible in the talk, partly unpacked how it is that the 'social sensitivity' found by Woolley et al operates within groups to lead to more effective thinking.

This paper proposes a concept, a simple Group Thinking Measure able, in combination with the analysis of videos of groups working together to solve a test, to offer a useful assessment of the quality of group thinking in classrooms. We have offered an initial proof of concept. The Group Thinking Measure is freely available and we invite others to further explore its potential.

References

Applebee, A. N., Langer, J. A., Nystrand, M., & Gamoran, A. (2003). Discussion-based approaches to developing understanding: Classroom instruction and student performance in middle and high school English. American Educational Research Journal, 40(3), 685-730.

Barad, K. (2007). Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning. Duke university Press.

Barnes, D., & Todd, F. (1977). Communication and learning in small groups. Routledge & Kegan Paul.

Barron, B. (2003). When smart groups fail. The journal of the learning sciences, 12(3), 307-359.

Davies, M., & Meissel, K. (2015). The use of Quality Talk to increase critical analytical speaking and writing of students in three secondary schools. British Educational Research Journal.

Dawes, L., Mercer, N., & Wegerif, R. (2000). Thinking Together: A Programme of Activities for Developing Thinking Skills at KS2. Questions Publishing Company.

De Oliveira, A. S., & Machado, A. L. (2015). Distance Education and Online Dialogues: Between Themes and Identities. Creative Education, 6(13), 1429.

EEF (2015) Philosophy for Children Evaluation report and Executive summary July 2015 Independent evaluators: Stephen Gorard, Nadia Siddiqui and Beng Huat See (Durham University) Accessed 15th July 2015 at https://educationendowmentfoundation.org.uk/uploads/pdf/Philosophy_for_Children.pdf

Engel, D., Woolley, A. W., Aggarwal, I., Chabris, C. F., Takahashi, M., Nemoto, K., .& Malone, T. W. (2015, April). Collective Intelligence in Computer-Mediated Collaboration Emerges in Different Contexts and Cultures. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (pp. 3769-3778). ACM.

Fernández, M., Wegerif, R., Mercer, N., & Rojas-Drummond, S. (2001). Re-conceptualizing " scaffolding" and the zone of proximal development in the context of symmetrical collaborative learning. The Journal of Classroom Interaction, 40-54.

Flecha, R. (2000). Sharing words: Theory and practice of dialogic learning. Rowman & Littlefield.

Forman, E. A., Larreamendy-Joerns, J., Stein, M. K., & Brown, C. A. (1998). “You're going to want to find out which and prove it”: Collective argumentation in a mathematics classroom. Learning and instruction, 8(6), 527-548.

Forman, E. A.,& Cazden, C. B. (1985). Exploring Vygotskian perspectives in education: The cognitive value of peer interaction. In J. V.Wertsch (Ed.), Culture, communication, and cognition: Vygotskian perspectives (pp. 323–347). New York: Cambridge University Press.

Habermas, J. (1979). What is universal pragmatics? In Communication and the Evolution of Society. Polity Press.

Howe, C. (2013). Scaffolding in context: Peer interaction and abstract learning. Learning, Culture and Social Interaction, 2(1), 3-10.

Howe, C., & Abedin, M. (2013). Classroom dialogue: A systematic review across four decades of research. Cambridge journal of education, 43(3), 325-356.

Hume, D. (1965/1751). An enquiry concerning human understanding. Alex Catalogue.

Kazak, S., Wegerif, R., & Fujita, T. (2015). The Kazak, S., Wegerif, R., & Fujita, T. (2015). The importance of dialogic processes to conceptual development in mathematics. Educational Studies in Mathematics, 90(2), 105-120.

Kincheloe, J. L. (2005). On to the next level: Continuing the conceptualization of the bricolage. Qualitative Inquiry, 11(3), 323-350.

Koschmann, T. D. (2011). Theories of learning and studies of instructional practice. New York, NY: Springer.

Kounios, J., and Beeman, M. (2009). The Aha! moment: The cognitive neuroscience of insight. Current Directions in Psychological Science, 18, 210–16

Kounios, J., & Beeman, M. (2015). *The Eureka Factor: Aha Moments, Creative Insight, and the Brain*. Random House.

Kuhn, D. (2015). Thinking together and alone. Educational Researcher, 0013189X15569530.

Kutnick, P., Ota, C., & Berdondini, L. (2008). Improving the effects of group working in classrooms with young school-aged children: Facilitating attainment, interaction and classroom activity. L earning and Instruction, 18(1), 83-95.

Matusov, E. (1996). Intersubjectivity without agreement. Mind, Culture, and Activity, 3(1), 25-45.

Matusov, E. (2011). Irreconcilable differences in Vygotsky's and Bakhtin's approaches to the social and the individual: An educational perspective. Culture & Psychology, 17(1), 99-119.

Mercer, N., Dawes, L., Wegerif, R., & Sams, C. (2004). Reasoning as a scientist: Ways of helping children to use language to learn science. British Educational Research Journal, 30(3), 359-377.

Mercer, N. (2000). Words and minds: How we use language to think together. Psychology Press.

Mercer, N. (2007). Sociocultural discourse analysis: Analysing classroom talk as a social mode of thinking. Journal of Applied Linguistics and Professional Practice, 1(2), 137-168.

Mercer, N., & Howe, C. (2012). Explaining the dialogic processes of teaching and learning: The value and potential of sociocultural theory. Learning, Culture and Social Interaction, 1(1), 12-21.

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

Mercer, N., Dawes, L., Wegerif, R., & Sams, C. (2004). Reasoning as a scientist: Ways of helping children to use language to learn science. British Educational Research Journal, 30(3), 359-377.

Mercer, N., Wegerif, R., & Dawes, L. (1999). Children's talk and the development of reasoning in the classroom. *British educational research journal*, *25*(1), 95-111.

Michaels, S., O'Connor, C., & Resnick, L. B. (2008). Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. Studies in philosophy and education, 27(4), 283-297.

Muukkonen, H., Lakkala, M., & Hakkarainen, K. (2009). Technology-Enhanced Progressive Inquiry in Higher Education. In M. Khosrow-Pour (Ed.), Encyclopedia of Information Science and Technology I-V. 2nd edition (pp. 3714-3720). Hershey, PA: Information Science Reference.

Neisser, U., Boodoo, G., Bouchard Jr, T. J., Boykin, A. W., Brody, N., Ceci, S. J., ... & Urbina, S. (1996). Intelligence: knowns and unknowns. American psychologist, 51(2), 77.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

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

Osberg, D. (2009). "Enlarging the space of the possible" around what it means to educate and be educated. Complicity: An International Journal of Complexity and Education, 6(1).

Resnick, L. B., & Schantz, F. (2015). Re-thinking Intelligence: schools that build the mind. European Journal of Education, 50(3), 340-349.

Reznitskaya, A., Kuo, L. J., Clark, A. M., Miller, B., Jadallah, M., Anderson, R. C., & Nguyen-Jahiel, K. (2009). Collaborative reasoning: A dialogic approach to group discussions. Cambridge Journal of Education, 39(1), 29-48.

Rojas-Drummond, S., Fernández, M., & Vélez, M. (2000). Habla exploratoria, razonamiento conjunto y solución de problemas en niños de primaria. In La Psicología Social en México, Vol. VIII, México: Asociación Mexicana de Psicología Social, 403-410.

Rojas-Drummond, S.M., Pérez, V., Vélez, M., Gómez, L., & Mendoza, A. (2003). Talking for reasoning among Mexican primary school children. *Learning and Instruction, 13 (*6), 1653-670

Sijtsma, K. (2009). On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika*, 74(1), 107–120. doi:10.1007/s11336-008-9101-0

Stahl, G. (2006). Group cognition: Computer support for building collaborative knowledge (pp. 451-473). Cambridge, MA: Mit Press.

Webb, N. M. (1989). Peer interaction and learning in small groups. International journal of Educational research, 13(1), 21-39.

Webb, P., & Treagust, D. F. (2006). Using exploratory talk to enhance problem-solving and reasoning skills in grade-7 science classrooms. Research in Science Education, 36(4), 381-401.

Wegerif, R. (1996) Using computers to help coach exploratory talk across the curriculum. Computers and Education 26 (1-3): 51-60. ISSN: 0360-1315

Wegerif, R. (2008). Dialogic or dialectic? The significance of ontological assumptions in research on educational dialogue. British Educational Research Journal, 34(3), 347-361.

Wegerif, R. (2011). Towards a dialogic theory of how children learn to think. Thinking Skills and Creativity, 6(3), 179-190.

Wegerif, R. (2013). Dialogic: Education for the Internet age. Routledge.

Wegerif, R., & Mercer, N. (1997). A dialogical framework for researching peer talk. Language and Education, 12, 49-64.

Wegerif, R., Mercer, N. & Dawes, L. (1999). From social interaction to individual reasoning: An empirical investigation of a possible sociocultural model of cognitive development. Learning and Instruction, 9(5), 493–516.

Wegerif, R., Perez Linares, J., Rojas-Drummond, S., Mercer, N., & Velez, M. (2005). Thinking together in the UK and Mexico: Transfer of an educational innovation. Journal of Classroom Interaction, 40(1), 40-48.

Wegerif, R., Perez, J., Rojas-Drummond, S., Mercer, N., & Velez, M. (2005) Thinking Together in the UK and Mexico: transfer of an educational innovation. Journal of Classroom Interaction, 40, 1, 40-48

Wertsch, J. V. (1988). Vygotsky and the social formation of mind. Harvard University Press.

White, E. J. (2015). Introducing dialogic pedagogy: Provocations for the early years. Routledge.

Woolley, A. Chabris, C. Pentland, A., Hashmi, N., & Malone, T. (2010) Evidence for a Collective Intelligence Factor in the Performance of Human Groups. Science, September 30, 2010 DOI:10.1126/science.1193147

Yang, Y. (2016). Lessons learnt from contextualising a UK teaching thinking program in a conventional Chinese classroom. Thinking Skills and Creativity, 19, 198-209.