Developing a test of reasoning for preadolescents

Ayesha Ahmed, Christine Howe, Louis Major, Sara Hennessy, Neil Mercer and Paul Warwick

Faculty of Education, University of Cambridge, Cambridge, UK

ABSTRACT
As part of an investigation into the relationship between classroom dialogue and student outcome, a test of reasoning has been developed that is suitable for preadolescents (i.e. c.10–13-year-olds). Building on previous work but expanding this considerably, the test focuses upon four areas of reasoning: differentiation of facts from opinions, differentiation of reasons from conclusions, inference of implications, and evaluation of multiple reasons. This paper reports on the test’s design and development (including the repeated cycles of piloting and redrafting), highlighting the methodological challenges that were faced (e.g. from linguistic demands), and how these were addressed. It also outlines how the test was trialled in English schools, including its use in a paper-based version with 218 students, and in a digital version with 129 students. Patterns of student performance were comparable across the two samples and also provide evidence for acceptable test reliability and validity. Thus, in addition to spotlighting methodological issues of general relevance within educational research, the paper presents a test that successfully assesses key aspects of student reasoning. The test is the first of its kind to be designed for this age group, and is offered as freely available for future research.

Introduction
There is growing consensus not simply about the significance of classroom dialogue for student learning but also about the maximally beneficial forms (see, e.g. Resnick et al. 2015). In particular, many scholars agree that optimal patterns include open questions; extended responses (ideally involving reasoned justifications); cumulative building upon others’ contributions; articulation, discussion and evaluation of competing viewpoints (even if none or only one is actually correct); and gradual resolution of differences towards productive conclusions. Moreover, while empirical scrutiny has been slow to take place (Howe and Abedin 2013), supportive evidence is beginning to appear (e.g. O’Connor et al. 2015, Pauli and Reusser 2015, Alexander et al. 2017, Jay et al. 2017). The putatively optimal patterns of dialogue do seem to be associated with boosts to the subsequent academic performance of individual students.

However, the focus of recent evaluations has mostly been upon what classroom dialogue implies for curriculum mastery. Important though this is, an additional issue is the implications for reasoning in a more general sense, not least because of suggestions that progress here mediates the relation between dialogue and curriculum mastery (Larrain 2013, now detailed in Larrain et al. 2020).
It certainly seems plausible that interaction around reasoned justification, appraisal of differences, and inferring conclusions should support facility with parallel aspects of reasoning at the individual level, yet the evidence is patchy. While the initiative known as Philosophy for Children emphasizes dialogue and is intended to support reasoning, evaluations have focused once more on curriculum mastery and have not hitherto provided direct evidence about dialogue-outcome relations (Trickey and Topping 2004, Topping and Trickey 2007). During the Thinking Together programme (Mercer et al. 1999, Mercer and Littleton 2007), hypothetically productive forms of dialogue were linked with scores on Raven’s Progressive Matrices (Raven and Raven 2003), a standardized test of nonverbal reasoning. However, implications for providing justifications, making inferences and drawing conclusions were not explored. Finally, associations have been demonstrated between the quality of reasoned justification during small-group interaction and their quality within the subsequent writing of individual students (Anderson et al. 2001, Fung and Howe 2016). Group work is, however, only one form of classroom interaction, and in any case the results are limited to a single aspect of reasoning.

Recognizing that evidence is currently limited, one aim of a project that the authors have recently completed was to explore dialogue-reasoning relations in depth (Howe et al. 2019). The project was funded by the UK’s Economic and Social Research Council (and is referred to hereafter as the ‘ESRC project’) and involved videotaping and coding the teacher-led dialogue occurring during English, Mathematics, and Science lessons. The focus was on Year 6 classrooms in England (10- to 11-year-old students, the oldest age band in English primary/elementary schools). Drawing on Hennessy et al. (2016), a scheme for coding was developed that reflects current views (both consensual and disputed) about productive forms of classroom dialogue. Project objectives involved relating variation in use of these forms to the subsequent curriculum mastery of individual students, their attitudes to schooling, and their skills in reasoning. In relation to the latter, the interest was specifically in the aspects of reasoning that have close semantic parallels with the putatively productive forms of dialogue, that is concepts like inference, reason and conclusion as spotlighted above together with the distinction between fact and opinion on which the other concepts in a sense depend. Given that the plan was to work with a large sample of students, it was hoped that an easy-to-administer instrument could be identified that would allow facility with these aspects of reasoning to be assessed through individual written responses in a classroom context.

A literature review indicated that currently available instruments for assessing reasoning skills fall into two main categories. The first category encompasses instruments designed to measure a construct variously called ‘aptitude’, ‘potential’, or ‘intelligence’. The afore-mentioned Raven’s Progressive Matrices is one example; others include the Cognitive Abilities Tests (CAT 4th Edition, G.L. Assessment 2016), and the Wechsler Intelligence Scale for Children (WISC-V, Wechsler 2016). Multiple-choice items in such tests are designed to measure abilities in cognitive domains including verbal, non-verbal, spatial and numerical reasoning, and provide separate indices or an aggregated score such as IQ. While such broad cognitive abilities may be related to an understanding of inferences, reasons, conclusions and facts versus opinions, the extent of the relation remains an open question (Haladyna and Rodriguez 2013, Newton and Shaw 2014). Thus, employing cognitive ability tests as proxies seemed inadequate given the ESRC project’s aims, and indeed a possible threat to validity (where validity is defined as ‘… the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses’: (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014, 21). Here the potential mismatch between the tests under consideration and the project’s aims suggested significant threats to validity through, in particular, ‘construct under-representation’ (Messick 1989).

The second category covers instruments designed to measure ‘critical thinking’. For instance, the California Critical Thinking Skills Test (CCTST, Facione 1991) addresses reflective decision-making using scenarios in which test-takers interpret texts or pictures and draw inferences. The Cornell Critical Thinking Test (Ennis et al. 1985) measures skills that include induction, deduction, and identifying assumptions. The Watson-Glaser Test addresses the ability to recognize assumptions, evaluate arguments and draw conclusions, and is often used in professional recruitment (Watson and Glaser 2002).
Moving away from the largely multiple-choice nature of these instruments, concept mapping has also been used to assess critical thinking. One example is *Thinking Maps* (Hyerle and Williams 2010) where students draw maps of content areas. Assessors then score the maps across eight dimensions that are based on a Piagetian conceptualization of reasoning. Collectively, the instruments that address critical thinking seemed closer to providing valid measures of the reasoning skills of interest. However, the available multiple-choice tests are designed for students considerably older than the ESRC project’s intended sample. They are also produced by commercial organizations and available only for substantial fees. As for concept mapping, a major drawback was the extensive training required for students to learn how to make maps and for assessors to learn how to score them.

Thus, no instrument emerged from the review that fully addressed the ESRC project’s requirements, and it became clear that something new would have to be designed. At the same time, assessments that had been prepared mainly for research purposes were identified that did include specific items of relevance. Past papers of the Oxford, Cambridge and RCA (OCR) Level 2 award in *Thinking and Reasoning Skills* (OCR 2009) contained items that tap the ability to differentiate facts from opinions. Burke and Williams’ (2012) thinking skills assessments address the distinction between reasons and conclusions. Yeh et al.’s (2000) ‘test of critical-thinking skills’ (TCTS-PS), developed in Taiwan and used also in Hong Kong (Fung and Howe 2016), contains items with good discriminant validity which assess recognition of implications and evaluation of multiple reasons. Although separately too specific for the ESRC project’s purposes, these tests offered a pool of items that seemed to provide foundations that could be built upon. Accordingly, the authors have been developing a new reasoning test that, although wider ranging than its predecessors, is grounded in this earlier work. In the sections that follow, the iterative and rigorous design work that underpinned the test’s development is outlined, with key methodological choices discussed in light of both the pre-existing items and the general literature on test development. Data are then presented that bear on the test’s reliability and validity. The over-arching aim throughout is to spotlight the strategies followed when addressing the challenges faced during test development for these are believed to be of general interest. At the same time, the hope is also to present an instrument that could prove valuable for other educational researchers.

**Methods of test development**

**Antecedents for the test**

Coverage of the key concepts in the most directly relevant literature suggested that they should be addressed via test items relating to four specific themes: the differentiation of facts from opinions, the differentiation of reasons from conclusions, the inference of implications, and the evaluation of multiple reasons. Thus, it was decided that test design should be structured around these four themes. Throughout the design process, a need was recognized: (a) to keep the conceptual framework firmly in mind, in keeping with principled approaches to assessment (Ferrara et al. 2016); (b) to obtain a chain of evidence that warrants inferences drawn from the test, as detailed in Kane’s (2001) argument-based approach to validity. In other words, it was seen as critical that item design be informed by a clear conception of the relevant aspects of reasoning, so that validity could be built into the test from the outset and well-founded claims could be made about the focal constructs.

Lane et al. (2016) emphasize the importance of a framework for collecting validity evidence, covering the intended constructs, population, test users and anticipated interpretations. A key component of this is obtaining support for the use of specific items. Accordingly, an iterative process of drafting, piloting and revision was applied in designing test items. In doing this, two further and more specific challenges had to be faced. First, because the subject matter was inherently linguistic there were dangers of construct-irrelevant variance from unnecessarily complex language
Minimizing this risk is essential for ensuring validity (see Haladyna and Rodriguez 2013, Abedi 2016). This latter warning bears also on the second specific challenge, namely the extent to which items should be located in contexts familiar to students. While there were likely advantages from using such contexts, there were also risks from, for example, reinforcing stereotypes (Zieky 2006). Moreover, it appeared unlikely that all of the four themes to be covered in the test could be addressed using familiar scenarios.

Mindful throughout of such issues, design work revolved around developing a four-section test, corresponding to the four key themes. In Draft 1, the first section (entitled Facts and Opinions and adapted from past papers of the OCR Level 2 award) comprised:

(a) a two-part item where Facts or Opinions were to be circled, e.g. ‘Points of view about something are called …’;
(b) a four-part item where statements like ‘The weather in Britain is awful’ were to be classified via ticks on a grid as Facts or Opinions;
(c) a three-part item where a character asserted that ‘Eating chocolate is a very unhealthy thing’, with the task being to indicate via ticks on a grid whether statements like ‘Champion runners can eat up to seven bars of chocolate a day’ support, challenge or are irrelevant to the assertion.

The initial three-part item in the second section (entitled Reasons and Conclusions) asked whether a character, as e.g. in ‘My dad asks me to find his tie’, was or was not looking for reasons. The second two-part item presented two short texts, with a reason to be underlined in the first text and a conclusion to be underlined in the second. Thereafter came four items (illustrated via the appendix, Q4), each containing a pair of texts. All texts were sourced from Burke and Williams (2012) but were adapted to minimize word length and ensure that texts in each pair were of similar length. The task was to identify the text that gave reasons and a conclusion, and to underline a reason (two items) or the conclusion (two items).

The third and fourth sections within Draft 1 were both inspired by material in Yeh et al.’s (2000) TCTS-PS, but since the TCTS-PS is heavily skewed towards the Chinese context, the material was substantially modified. Specifically, Draft 1’s third section (entitled Saying and Implying) began by explaining the distinction between saying and implying, and via one two-part item checking that the distinction was understood. Thereafter came six items (some adapted from Yeh et al. and some newly designed) where an implication of a short text had to be selected from three options (see the appendix, Q6). The fourth section (entitled Spotting Good Reasoning) contained five items, which all started with one character expressing an opinion, e.g. ‘Ms Jones says all primary school children in England should learn computing’. Each continued via speech bubbles depicting two further characters giving reasons for agreeing or disagreeing, e.g. one character said ‘I agree because technology is the future. If children don’t learn computing, they won’t be properly prepared. This will cause problems for society’. Another character then replied ‘I disagree because primary school children play games on computers, which will affect their schoolwork and hurt their eyes’. The task was to indicate which of the latter two characters gave the best reasons (or whether the quality did not differ), and write a sentence justifying the choice.

**Pilot 1**

Draft 1 was presented in a single classroom to students aged 10–11 years (see Table 1), with responses analyzed via Cronbach’s alpha and Rasch modelling (Cronbach 1951, Rasch 1960). Given the small sample size, the results reported in Table 1 were regarded as no more than indicative. Nevertheless, they were encouragingly compatible with internal scale consistency (Cronbach) and with uni-dimensionality amongst test items (Rasch out-fit z-std, Bond and Fox 2007). Yet
while there was no reason to doubt the quantitative message with Sections 1, 2 and 3, the justifications written for Section 4 (the hardest section as judged through the proportion of correct answers) highlighted a problem. In many cases, the students first decided whether or not they agreed with the initial claim, and then chose the character concurring with their decision. For instance, if they believed technology should be taught in primary schools, they would select the character endorsing this view. There was no apparent scrutiny of the character’s reasons, yet the point of the item was to assess the evaluation of reasons. Moreover, on the rare occasions that reasons were referred to, the students’ justifications for selecting what were initially regarded as wrong answers were sometimes so persuasive that the original interpretations began to seem doubtful. Consideration was given to dropping Section 4 altogether, but the distribution of scores for Sections 1, 2 and 3 suggested that these sections alone would make the test too easy. Moreover, the aspect of reasoning that Section 4 was intended to tap was clearly important, suggesting that coverage should be retained if at all possible. The next step therefore was to examine whether evaluation of multiple reasons could be approached with valid but still challenging items, and therefore preserved in the test.

Pilot 2

Drafts 2a and 2b and their respective pilots (see Table 1) were designed to explore alternative approaches to Section 4, adopting a highly focused approach but with very small samples. One major conclusion from Draft 2a was that designing items around three characters (the initial claimant and two respondents) made it difficult to clarify which character was referred to and caused unnecessary confusion, i.e. construct-irrelevant variance. It was preferable to restrict items to two characters, e.g. one saying ‘I think all primary school children should learn computing because technology is the future’… ‘ and the other disagreeing. A second major conclusion was that focusing upon the content of reasons was not productive: if the content differed markedly in quality all students succeeded in identifying the character who gave the best reason, whereas if the quality was similar, responses continued, as with Pilot 1, to be based on which claims the students subscribed to. The characters’ reasons were not used in evaluation. On the other hand, meaningful discrimination amongst students could be made if the focus shifted from content to discursive adequacy. Putting

| Pilot | Sample | Procedure | Results |
|-------|--------|-----------|---------|
| 1     | N=29; age=10–11 years; full Year 6 class from village primary school | Booklets presenting Sections 1–4 (Draft 1) completed as timed (45 min) test; afterwards students drew smiley, neutral, or sad faces to indicate whether each section was easy, moderate or hard | Cronbach’s alpha=0.71; 34/35 items with out-fit z-std between +2 and −2 on Rasch analysis; modal evaluation of Sections 1 & 3 = easy, and of Sections 2 & 4 = moderate |
| 2a    | N=8; age=11–12 years; top and lower maths sets from Year 7 of village secondary school | Booklets presenting Section 4 in two formats (Draft 2a) completed without time restrictions; afterwards each maths set separately interviewed | Cronbach and Rasch not applicable; problems detected with: (1) keeping track of 3 characters; (2) evaluating content of reasons |
| 2b    | N=14; age=11–12 years; from Year 7 of village secondary school; teacher rated students as ‘mixed ability’ | Booklets presenting Section 4 prepared on basis of Study 2a results (Draft 2b); administered as for 2a; follow-up discussion with whole group, and individual rating of items on 1 (easy) to 10 (hard) scale | Cronbach and Rasch not applicable; difficulty ratings all between 2 and 5, yet only 36% success; confirmed value of: (1) reducing text to 2 characters; (2) focusing on discursive adequacy |
| 3     | N=24; age 10–11 years; full Year 6 class of ethnically diverse, low SES London primary school | Booklet presenting Sections 1–4 (Draft 3) completed as for Study 1; feedback session with whole class, and later with teacher | All students completed in 25–35 min; Cronbach’s alpha=0.75; 43/45 items with out-fit z-std between +2 and −2 on Rasch analysis; scores ranged from 14 to 37 with normal distribution; class claimed to enjoy test and find it easy |

1All schools located in SE England, and with records of national test performance slightly above national averages.
these two conclusions together, items were designed where one character made a reasoned claim and the other: (a) ignored the reason (discursively inadequate), e.g. ‘The main problem is that they [ex-prisoners who return to crime] can’t get a job after they’re released’ and ‘I agree because mixing with the wrong crowd could lead them back to crime’; or (b) repeated the reason (also discursively inadequate), e.g. ‘… when children misbehave, it [keeping them in at lunch-time] can sometimes be the only answer’ and ‘I agree because keeping the children in at lunch-time can sometimes be the only way to stop them misbehaving; or (c) engaged with the reason and supplemented it (discursively adequate), e.g. ‘I think that not doing homework is one reason why children do badly in school’ and ‘But in some countries children don’t get homework, and they do just as well as the kids here’. The task was to identify the relationship between reasons from options, i.e. whether the second character’s reason ‘doesn’t really relate to’ what was said, ‘isn’t really different’, or offers ‘a new reason to think about’. Problem content was drawn from Kuhn (1991), Weisberg et al. (2008), and Yeh et al. (2000).

Pilot 3

A four-section Draft 3 was prepared and administered to a class of 10- to 11-year-olds. Sections 1, 2 and 3 were largely as in Draft 1 apart from minor changes to wording. The only substantive adjustment was to re-order Section 3, so that items were sequenced in order of difficulty as judged from Pilot 1’s Rasch modelling. Section 4 (now entitled Talking about Reasons) started with two introductory items where one character (Susie) made a claim and gave a reason, and a second character (Jamie) replied giving a reason. The task in both cases was to underline the reason. Then came an item where Susie was said to reflect on how Jamie’s reason related to hers, and the task was to choose the option that expressed the relation. Six pairs of items with the same format followed. Draft 3 was piloted and while the sample was once more too small for conclusive results, the quantitative data were encouraging. As indicated in Table 1, the Rasch out-fit z-std of two Section 3 items lay outside the acceptable range of +2 to −2 (despite having been within-range in Pilot 1), and these items were reworded. With the Section 4 items, the values for out-fit z-std were never larger than 1.4. Yet several students asked for help with reading Section 4’s instructions, and when the test was discussed subsequently with the class and their teacher, instruction wordiness was mentioned. Moreover, two pairs of items were felt to contain ambiguities. Thus, these two pairs were omitted from Section 4, and task instructions relating to the remaining items were simplified. In particular, detail was phased out across items once grasp of requirements could be safely presumed. The appendix presents part of an instruction-simplified item from Section 4 (Q7a), highlighting the two-character exchange of reasons and what amounted to five response options, three options covering the relations between reasons and two ‘distractors’ that reiterated the claim or introduced a new idea.

Results from test trialling

With the minor adjustments made after Pilot 3, the reasoning test now comprised:

(a) Section 1: Facts and Opinions – as with Drafts 1 and 3, apart from making one item three- rather than four-part (to require 40 responses in the overall test rather than 41);
(b) Section 2: Reasons and Conclusions – as with Drafts 1 and 3;
(c) Section 3: Saying and Implying – as with Draft 3, apart from minor changes of wording;
(d) Section 4: Talking about Reasons – the introductory Susie/Jamie item as with Draft 3, and four pairs of items as in the appendix.

Regarded as ready for trialling with a larger sample, this version was administered in the first nine Year 6 classes (students aged 10–11 years) recruited for the ESRC project (N=218 students). This time,
the number of participants was more than adequate for relevant quantitative analyses (Bond and Fox 2007). The selected classes were in rural and urban locations across SE England (including London) and were socio-economically and ethnically diverse ($M=31.66\%$ of students in each class eligible for free school meals, range=$3.57\%$ to $65.38\%$; $M=42.09\%$ from minority ethnic backgrounds, range=$0\%$ to $92.59\%$). While English was the second language for some students, all teachers judged $100\%$ of their class to be fluent in English. The tests were printed in booklets and administered by teachers after receipt of detailed written procedural guidance. They were completed under examination conditions, with the students allowed 30 min (reduced from Pilots 1 and 3 due to fewer items).

The sample mean score on the test was 22.74 with a standard deviation of 4.82 and a score range from 11 to 35 (out of a possible 40). Internal consistency as measured by Cronbach’s alpha was very high at 0.98. Rasch modelling indicated that with 38 out of the 40 items, the out-fit z-std values were between $+1.8$ and $−1.9$, with a mean of $0.1$ and standard deviation of $1.1$. Of the two items that were outside the conventionally desirable range, one showed misfit, with an out-fit z-std of $2.7$. This was a relatively difficult item from Section 4. However, there were several similar items in the test that did not show misfit, so a decision was taken not to discard this item. The other item falling outside the desirable range showed over-fit (z-std $= −2.5$). It was a comparatively easy item from Section 2, and there was no obvious reason to discard it, especially when the associated out-fit z-std had been very small in both Pilot 1 and Pilot 3. These two items apart, the results strongly suggested that a uni-dimensional construct was being measured, implying a degree of coherence across mastery of the concepts being assessed.

One major advantage of Rasch modelling over, say, factor analysis is that the former provides information about the score distribution of both student respondents and test items (see again Bond and Fox 2007). Comparison across the distributions obtained from the trial indicated that mean item difficulty was broadly in line with mean student performance on the test, indicating that the items were appropriately targeted for this age level. The initial (Section 1) items, checking knowledge of the definitions of ‘facts’ and ‘opinions’, were relatively easy and were designed for students to perform well at the start. The most difficult item was in Section 2 and required students to underline a conclusion. In this particular item, the conclusion ‘Teaching is the best job in the world’ was actually given in the first sentence of the paragraph, followed by reasons. This is an unusual construction as conclusions typically lie at the end of passages. The difficulty of this item was not therefore especially surprising. However, the item showed no evidence of misfit (out-fit z-std of $0.4$), meaning that those students who did better on this item did better throughout the test, and those who found it most difficult were low scorers on the test as a whole. It can therefore be concluded that the item’s difficulty is valid in relation to what was being measured.

Regarding the ESRC sample’s results as promising, an adapted version of the reasoning test was subsequently used in a ‘Digital Dialogue’ project funded by the Research Council of Norway (FINNUT/Project No: 254761). Involving schools based in England and Norway, and in this case including students aged 11–13 years, this project investigates the role of using a micro-blogging tool in supporting, enhancing and modifying classroom dialogue. Due to a need for rapid test distribution across two countries, and the digital nature of the micro-blogging tool itself, the research team was keen to develop a digital version of the reasoning test to explore the development of students’ reasoning ability. Such a version was expected to make limited time and resource demands, in terms of both administration (as email distribution is cost-effective and quick) and analysis (as simple procedures such as frequency counts could be automated).

Inevitably, minor modifications had to be made for ‘on-screen’ administration. Several items in the ESRC project’s paper-based version involved circling or underlining phrases or sentences; these were changed to mouse clicks or typing key words. For instance, instead of requesting that a sentence be underlined, the digital version of the first item in the appendix (Q4) required respondents to ‘Type the first two words of a sentence from one of the texts that gives a reason’. Motivated by the need to accommodate a range of screen sizes (desktop computers, tablets etc.), Section 4 items were divided into two parts, with students clicking a ‘forward’ arrow to reach Part
B. Similarly, while character images were used to illustrate Section 4 items on paper, these were removed in the digital version to ensure consistency of presentation across different devices. After initial piloting in England and (after translation) Norway, the reasoning test was implemented in its digital form (using the Qualtrics platform). Data from two secondary schools in the East of England were analyzed with a view to comparison with the ESRC sample. This sample comprised 129 students from Year 7, the first year of secondary education (age range 11–12 years).

The mean score for the Digital Dialogue sample was 25.86 with a standard deviation of 4.19 and a range from 15 to 38 (out of a possible 40). Thus, performance was similar to that of the ESRC sample, but slightly better. The latter is unsurprising when the Digital Dialogue sample came from students who were a year older. As Table 2 shows, the two samples also performed very similarly across Section 4, where the items in the paper and digital versions diverged most markedly. Excluding the introductory ‘Susie and Jamie’ items (Q1 to Q3), Table 2 shows the percentage of correct answers with each Section 4 item. Both the absolute level of difficulty and the rank order were very similar across the two versions. For both versions, the most and least difficult items were the same. In fact, only two items in Section 4 were ranked differently in order of difficulty.

Discussion and conclusions

Reflecting an interest in the interaction between aspects of reasoning and classroom dialogue, the aim was to develop a test of reasoning for preadolescents (approximately 10–13 years of age). Of particular interest was understanding of the differences between facts and opinions, between reasons and conclusions, and between saying and implying together with the ability to evaluate relationships between multiple reasons. Having reviewed existing instruments and finding that none were sufficient for the intended purposes, a new test was developed to assess these four areas. Development has involved adapting items from existing tests and writing new items. The key challenge was ensuring that the sets of items support valid inferences relating to the aspects of reasoning of interest. To this end a rigorous, iterative process of piloting and review was implemented, with particular attention paid to the interaction between the cognitive and linguistic demands of the items. Two trials have also been conducted with large samples of students.

As noted, trialling of the paper-based version during the ESRC project indicated that the test was highly reliable (Cronbach’s alpha = 0.98). Furthermore, Rasch modelling demonstrated a reasonable approximation to uni-dimensionality, with 38 out of 40 items achieving out-fit z-std values between +1.8 and −1.9. There is of course no a priori necessity to strong links amongst understanding of facts, opinions, inferences, reasons and conclusions, so the psychometric simplicity indicated by the Rasch modelling is both informative and reassuring. Further trialling with the digital version produced a comparable spread of scores to the paper-based version, both overall and after item-by-item comparison across Section 4. In general, then, the test fares well against standard psychometric indicators, suggesting that at the very least it constitutes a reliable measure with the age group in question.

| Table 2. Percentage of Sample Succeeding with Section 4 Items. |
|----------------------|----------------------|
|                      | ESRC Sample     | Digital Dialogue Sample |
| Q4a                  | 25.69%          | 31.78%               |
| Q4b                  | 44.95%          | 40.31%               |
| Q5a                  | 10.09%          | 8.53%                |
| Q5b                  | 31.19%          | 37.98%               |
| Q6a                  | 28.44%          | 21.71%               |
| Q6b                  | 16.97%          | 16.27%               |
| Q7a                  | 7.34%           | 6.98%                |
| Q7b                  | 22.48%          | 33.33%               |
However, did the reliable test also measure what it was intended to measure, i.e. reasoning skills across the four targeted themes? In other words, was the test valid for its intended purpose? The importance of validity was kept in mind throughout the development process. Following Kane (2001), this meant being aware of the inferences to be drawn from test scores, and therefore the evidence needed for those inferences to be warranted. In practical terms, it was important that item design be grounded in a clear conceptual framework, drawing on literature and being mindful of limitations in existing instruments. Both reliability and validity were maximized through close collaboration within the research team at each stage, including in-depth discussions of perceptions of items.

Arguably, the greatest methodological challenge was designing items that minimized construct-irrelevant variance caused by language demands, a well-documented hurdle when test-takers are children (Pollitt et al. 2007, El Masri et al. 2016). The challenge was particularly marked with the present test, for the inherently linguistic nature of the subject matter meant that language was a source of construct-relevant variance as well as construct-irrelevance and the line between the two was not clear-cut. While there was always a balance to be struck between simplification and clarity, an attempt was made throughout to minimize text as much as possible and to use complex language only when it was an essential part of the item (Abedi 2016). Strategies like reducing text across equivalent items will also have helped: for instance, detail provided early in Section 4 allowed instructions to be compressed in the final item to, e.g. ‘The professor tells Peter correctly how the reason he gave relates to what she said’ (see the appendix). Nevertheless, while it is hoped that superfluous language demands were generally avoided, it is also recognized that the test may, of necessity, have been linguistically taxing for some students in this age group.

Of course, when many test items were derived from pre-existing instruments, prior usage provides grounds for confidence in validity despite possible linguistic limitations. Moreover, results obtained subsequently with the test itself are also encouraging. In the ESRC project, the reasoning test was eventually completed by 1766 10- to 11-year-olds from 72 ethnically and socio-economically diverse Year 6 classrooms spread across England (Howe et al. 2019). Test scores were related to specific indices of teacher-led dialogue in accordance with the project’s main aim, but group work amongst students was also assessed. While results relating to teacher-led dialogue are largely beyond the present paper’s scope, the relation between quality of group work and test scores is relevant. Quality was assessed via ten 3-point scales, with five scales addressing such general features as whether all students were involved, and five scales addressing group dialogue including, crucially, justification, disagreement, and elaboration. Thus, the features that Anderson et al. (2001) and Fung and Howe (2016) found to be predictive of reasoned writing were prominent amongst the quality indicators. Therefore, it seems highly encouraging as regards the test’s validity that aggregated group work ratings with the ESRC sample predicted reasoning test scores, \( t(72.88) = 2.53, p = .014 \), with prior attainment taken into account. Moreover, this result remained significant even when indices of teacher-student dialogue were factored in.

Given the promising pilot results in the Digital Dialogue project, the reasoning test in its digital form was used with a further 222 students in Norway. These students were aged 13–14 years and drawn from seven classes across four secondary schools in the Oslo region. The test was completed before \( (M=24.32, SD=5.24) \) and after \( (M=26.01, SD=5.21) \) the introduction of a dialogic pedagogy supported via the micro-blogging tool. Differences between mean before and after scores proved statistically significant, \( p<0.01 \). Students scoring in the lowest quartile before introduction of the dialogic pedagogy progressed the most, a trend that was particularly strong for those classes that showed the most progress overall. For example, in one class, the students who scored in the lowest quartile before introduction scored 7 points higher afterwards. Further analysis is being undertaken to interrogate these preliminary findings, but they are exactly what would be expected after a dialogically based intervention in which reasoning necessarily played an integral part. Therefore, they strengthen faith in the test’s fundamental validity.
Further trialling would be welcome, and for that reason the test is being made freely available within the research community (see Supplementary Material). This said, it is important to stress that in order to preserve validity the test, like all others, should only be used for its intended purposes, for otherwise claims made from scores cannot be justified. The point is perhaps especially relevant for tests designed to assess reasoning skills at the stage of schooling that the ESRC and Digital Dialogue projects address (transition between primary and secondary schooling). At this stage in England at least, students take compulsory tests that have had a history of being used for multiple purposes, some of which are at odds with their original intention (Newton 2007). Specifically, there are pressures on schools to assess children’s ‘ability’ or ‘potential’ due to high stakes accountability systems (see Torrance 2018). The aim for the reasoning test reported here is that it be used exclusively for research purposes.

**Disclosure statement**

No potential conflict of interest was reported by the author(s).

**Funding**

This work was supported by the Economic and Social Research Council [grant number ES/M007103/1].

**ORCID**

Ayesha Ahmed http://orcid.org/0000-0003-1099-788X  
Christine Howe http://orcid.org/0000-0002-5383-4784  
Louis Major http://orcid.org/0000-0002-7658-1417  
Sara Hennessy http://orcid.org/0000-0002-9050-4995  
Neil Mercer http://orcid.org/0000-0002-6829-8072  
Paul Warwick http://orcid.org/0000-0002-0463-150X

**References**

Abedi, J., 2016. Language issues in item development. In: S. Lane, M. Raymond, and T. Haladyana, eds. *Handbook of test development*, 2nd ed. New York: Routledge, 355–373.

Alexander, R., Hardman, F., and Hardman, J., 2017. Changing talk, changing thinking. interim report from the in-house evaluation of the CPRT/UoY dialogic teaching project. www.robinalexander.org.uk.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014. *The standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anderson, A., et al., 2001. Peer interaction and the learning of critical thinking skills in further education students. *Instructional science*, 29, 1–32. doi:10.1023/A:1026471702353

Bond, T. and Fox, C., 2007. *Applying the Rasch model: fundamental measurement in the human sciences*. 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates.

Burke, L. and Williams, J., 2012. Two thinking skills assessment approaches: ‘assessment of pupils’ thinking skills’ and ‘individual thinking skills assessments’. *Thinking skills and creativity*, 7, 62–68. doi:10.1016/j.tsc.2011.11.002

Cronbach, L., 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334. doi:10.1007/bf02310555

El Masri, Y., Baird, J.-A., and Graesser, A., 2016. Language effects in international testing: the case of PISA 2006 science items. *Assessment in education: principles, policy & practice*, 23, 427–455. doi:10.1080/0969954X.2016.1218323

Ennis, R., Millman, J., and Tomko, T., 1985. *Cornell critical thinking tests level X and level Z: manual*. Pacific Grove, CA: Midwest Publications.

Facione, P., 1991. *Using the California critical thinking skills test in research, evaluation, and assessment*. Millbrae, CA: Academic Press.

Ferrara, S., et al., 2016. Principled approaches to assessment design: development and implementation. In: A. Rupp and J. Leighton, eds. *The Wiley handbook of cognition and assessment: frameworks, methodologies and applications*. Chichester: Wiley Blackwell, 41–74.
Fung, D. and Howe, C., 2014. Group work and the learning of critical thinking in the Hong Kong secondary liberal studies curriculum. *Cambridge journal of education*, 44, 245–270. doi:10.1080/0305764X.2014.897685

G.L. Assessment, 2016. *Cognitive abilities test*. 4th ed. London: GL Assessment.

Haladyna, T. and Rodriguez, M., 2013. *Developing and validating test items*. New York: Routledge.

Howe, C., Hennessy, S., G.L. Assessment, 2016. Classroom dialogue: a systematic review across four decades of research. *Cambridge journal of education*, 43, 325–356. doi:10.1080/0305764X.2013.786024

Hyerle, D. and Williams, K., 2010. Bifocal assessment in the cognitive age: thinking maps for assessing content learning and cognitive processes. *New hampshire journal of education*. Available from: https://www.thinkingfoundation.org/whole-school-change

Jay, T., et al., 2017. *Dialogic teaching: evaluation report and executive summary. project report*. London, UK: Education Endowment Foundation.

Kane, M., 2001. Current concerns in validity theory. *Journal of educational measurement*, 38, 319–342. doi:10.1111/j.1745-3984.2001.tb01130.x

Kuhn, D., 1991. *The skills of argument*. Cambridge: Cambridge University Press.

Larrain, A., 2013. Argumentation in science learning in chilean primary education: an efficacy study. Paper presented at the 15th biennial EARLI and JURE conference, Munich, Germany, August 27–31.

Larrain, A., et al., 2020. Argumentation skills mediate the effect of peer argumentation on content knowledge in middle-school students. *Journal of educational psychology*. Advance online publication. doi:10.1037/edu0000619

Mercer, N. and Littleton, K., 2007. *Dialogue and the development of children’s thinking: a sociocultural approach*. London: Routledge.

Mercer, N., Wegerif, R., and Dawes, L., 1999. Children’s talk and the development of reasoning in the classroom. *British educational research journal*, 25, 95–111. doi:10.1080/0141192990250107

Messick, S., 1989. Meaning and values in test validation: the science and ethics of assessment. *Educational researcher*, 18, 5–11. doi:10.2307/1175249

Newton, P., 2007. Clarifying the purposes of educational assessment. *Assessment in education: principles, policy & practice*, 14, 149–170. doi:10.1080/09695940701478321

Newton, P. and Shaw, S., 2014. *Validity in educational and psychological assessment*. London: SAGE Publications.

O’Connor, C., Michaels, S., and Chapin, S., 2015. ‘Scaling down’ to explore the role of talk in learning: from district intervention to controlled classroom study. In: L. Resnick, C. Asterhan, and S. Clarke, eds. *Socializing intelligence through academic talk and dialogue*. Washington, DC: American Educational Research Association, 111–126.

Oxford, Cambridge, and RCA (OCR), 2009. *Thinking and reasoning skills level 2 award*. Available from: https://www.ocr.org.uk/qualifications/other-general-qualifications/other-general-qualifications-thinking-and-reasoning-skills-level-2-award-j930/

Pauli, C. and Reusser, K., 2015. Discursive cultures of learning in (everyday) mathematics teaching: a video-based study on mathematics teaching in German and Swiss classrooms. In: L. Resnick, C. Asterhan, and S. Clarke, eds. *Socializing intelligence through academic talk and dialogue*. Washington, DC: American Educational Research Association, 181–193.

Pollitt, A., Ahmed, A., and Crisp, V., 2007. The demands on examination syllabuses and question papers. In: P. Newton, J.-A. Baird, H. Goldstein, H. Patrick, and P. Tymms, eds. *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority, 166–206.

Rasch, G., 1960. *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut.

Raven, J. and Raven, J., 2003. Raven progressive matrices. In: S. McCallum, ed. *Handbook of nonverbal assessment*. Boston, MA: Springer, 223–237.

Resnick, L., Asterhan, C., and Clarke, S., 2015. *Socializing intelligence through academic talk and dialogue*. Washington, DC: American Educational Research Association.

Topping, K. and Trickey, S., 2007. Collaborative philosophical inquiry for schoolchildren: cognitive gains at 2-year follow-up. *British journal of educational psychology*, 77, 787–796. doi:10.1348/000709907X193032

Torrance, H., 2018. The return to final paper examining in English national curriculum assessment and school examinations: issues of validity, accountability and politics. *British journal of educational studies*, 66, 3–27. doi:10.1080/00071005.2017.1322683

Trickey, S. and Topping, K., 2004. ‘Philosophy for children’: a systematic review. *Research papers in education*, 19, 365–380. doi:10.1080/0267152042000248016

Watson, G. and Glaser, E., 2002. *Watson-Glaser critical thinking appraisal*. London: Pearson Assessment.

Wechsler, D., 2016. *Wechsler intelligence scale for children*. 5th ed. Bloomington, MN: Pearson.
Appendix: Sample Items from the Paper-based Reasoning Test

Section 2 (Q4)
Read Box 1 and Box 2.

Box 1
An orchestra is made up of lots of different instruments. Some orchestras are very large and can have up to 100 players in them. The leader of the orchestra is normally called a conductor.

Box 2
There are many benefits of attending after-school clubs. You get to find out more about your teachers, there is a wide choice of activities, you make new friends and you learn new skills.

(a) Tick the box that gives reasons and a conclusion.
(b) Now underline a sentence that contains a reason.

Section 3 (Q6)
Anyone who wins at tennis must be good at it; most people who are good at tennis use good rackets.
What do these sentences imply?
Tick one of the boxes.

People who lose at tennis don’t use good rackets.
Most people who win at tennis use good rackets.
People who use good rackets must be good at tennis.
Section 4 (Q7a)
A professor is talking to her students.

She says:

One theory is that if you know something, you assume others know it too. Do you believe this theory?

Peter replies:

Yes, I believe it because brain scans show the front of the brain is involved in self-knowledge. People are good at judging themselves.

The professor tells Peter correctly how the reason he gave relates to what she said. She starts “I’m pleased you accept the theory …” What did the professor say next about how Peter’s reason relates? Tick ONE box to complete her comment.

A  “… and your reason gives some evidence for why the theory is correct”
B  “… because if you know something you do think others know it too”
C  “… but your reason isn’t really different from mine”
D  “… because if people trust others their answers might be right”
E  “… but your reason doesn’t really relate to the theory”