Int J of Sci and Math Educ
DOI 10.1007/s10763-017-9844-0

Progression Trend of Scientific Reasoning from Elementary School to University: a Large-Scale Cross-Grade Survey Among Chinese Students

Lin Ding[1]

Received: 16 December 2016 / Accepted: 30 July 2017

Abstract This study examines progression trends of Chinese students' scientific reasoning skills across grade levels from elementary school to university. A large-scale survey using the Classroom Test of Scientific Reasoning (CTSR) was conducted with 2669 Chinese students at 13 grade levels (grades 4–16). The construct validity of the CTSR was first examined using Rasch analysis to verify that the various reasoning sub-skills targeted by the test (hypothetical-deductive reasoning, proportional reasoning, correlation reasoning, probabilistic reasoning, and control of variables) form a cohesive cognitive construct. Based on an affirmative result, we proceeded to fit grade-level averages to the best-fit logistic regression model and generated trend lines for the reasoning sub-skills measured by the CTSR. Results show that the cross-grade progression trends generally follow continuous paths through increasingly greater levels of improvement. While all sub-skills display a noticeable growth during middle and high school years, they progress across the grades with different rates. Specifically, the development of proportional reasoning is fast-paced and large, but the progression of hypothetical-deductive reasoning and control of variables is markedly tardy and small. Formal science instruction, particularly the emergence of multiple domain-specific science courses in middle and high school, can influence the cross-grade progression trends. Possible causes and future improvement in science education are proposed. This study contributes to our understanding about the development of scientific reasoning and offers important implications for science instruction and research at different grades.

Keywords Classroom test of scientific reasoning · Cross-grade progression · Scientific reasoning · Survey

✉ Lin Ding
ding.65@osu.edu

[1] Department of Teaching and Learning, The Ohio State University, Columbus, OH 43210, USA

Published online: 19 August 2017

One of the major goals for science education is to promote students' scientific reasoning skills so that they can become more capable of making informed decisions and solving complex problems in their future workplace (National Research Council [NRC], 2000, 2012). In recent years, a growing body of literature has emerged, suggesting that advanced reasoning skills, such as generation and evaluation of evidence-based hypotheses, can have a positive impact on students' learning of science content and understanding of the nature of science (Ates & Cataloglu, 2007; BouJaoude, Salloum & Abd-El-Khalick, 2004; Ding, 2014; Kuhn, Black, Keselman & Kaplan, 2000; Lawson, 2004; Osborne, 2010; Schauble, 1996; Shayer & Adey, 1993). Researchers and educators therefore have invested much effort in designing effective instructional strategies to cultivate the development of student scientific reasoning in general science courses.

It is well received that instructional interventions on this matter must take into consideration student current reasoning skills (Coletta & Phillips, 2005; Coletta, Phillips & Steinert, 2007a, b; Ding, Wei & Liu, 2016; Ding, Wei & Mollohan, 2016; Meltzer, 2002). Since students at various grades possess different levels of such skills, successful design and enactment of interventions hinges critically on a deep understanding of how student reasoning progresses through grades. Previous studies on this topic, albeit extensive and useful, were scattered at different coarse-grained ages/stages and have yet systematically looked into a continuous trend across consecutive grades. A study capable of revealing a finer-grained progression trend of student reasoning across multiple grade levels is particularly useful, because it not only can deepen our understanding of this developmentally cognitive construct but also can guide effective design and implementation of grade-appropriate science curricula and instruction.

Besides the above limitation, prior studies in this arena also shared a common feature; that is, much attention was devoted to investigating scientific reasoning skills among students in Western countries. Results of such studies almost unanimously supported a mutually corroborative relationship between science content learning and scientific reasoning. However, a recent cross-nation study that compared American and Chinese university students in this matter pained a vastly different picture, showing that although Chinese students outperformed their American counterparts in physics content knowledge, their scientific reasoning skills demonstrated no significant advantage (Bao, Cai, Koenig, Fang, Han, Wang et al., 2009). This study leads to many interesting questions regarding the scientific reasoning skills of Chinese students and particularly the development theory of this topic in a non-Western educational system. For example, does the asynchrony between science content and scientific reasoning among Chinese college students suggest their reasoning skills will reach a plateau at the end of high school and therefore are no longer in phase with their content learning? Moreover, what further spurs an already growing interest in this student population is the leading performance of Chinese students on international assessments, such as the OECD Program for International Students Assessment (PISA). Collectively, these studies and reports have all increasingly heightened a need to initiate more extensive studies into Chinese students' scientific reasoning and learning. To this end, we seek to investigate the cross-grade progression of Chinese students' scientific reasoning in a broader context extending from elementary school to higher education institutions.

Typically, scientific reasoning skills can be measured through cognitive interviews or written tests. The former approach, such as those conducted by Kuhn (2002) and Schauble (1996), probes how students determine causal mechanisms in a complex

situation through repeated hands-on experimentations. This method can offer detailed, rich information and is particularly suitable for small-scale intensive studies. On the other hand, the paper-and-pencil approach to measuring student reasoning through a series of written tasks can yield general patterns of a large population more efficiently. Given the need for a large-scale survey of student reasoning across multiple grade levels, the written test approach using the Classroom Test of Scientific Reasoning (CTSR; Lawson, 1978) is carried out in this study (see the "Methods" section for more details). Because the CTSR has undergone several revisions since its initial release, we used the latest version (the 2000 version). Interestingly, despite its broad usage in many prior studies, the validity and reliability of this newly revised CTSR have heretofore only been evaluated through the classical test theory and hence are sample dependent. This study further examines these issues through Rasch analysis and explores the extent to which the individual items on the latest CTSR measure a common cognitive construct of scientific reasoning.

The structure of the remainder of the paper is organized as follows. We first review the key patterns and the development of scientific reasoning skills. Based on the developmental framework of this topic, we propose to study the cross-grade progression trend of scientific reasoning among Chinese students from elementary school to university. The methods and results of the study are reported, followed by discussions and implications.

What Are Scientific Reasoning Skills?

Scientific reasoning skills, according to Lawson (2004), are mental plans, strategies, or rules used to process information and derive conclusions beyond direct experiences. Chief among them is hypothetical-deductive reasoning, in which a person observes a puzzling phenomenon, generates plausible explanations (hypotheses), deduces inferences (predictions), and then plans and carries out experiments to test the predictions. By comparing the predictions against experimental outcomes, one can then decide to accept, reject, or revise the hypotheses. This hypothetical-deductive process is considered to be the core of scientific reasoning. Historically, scientists relied heavily on this reasoning process to build causal models and theories to make sense of the natural world (Lawson, 2000, 2005; Lawson, Oehrtman & Jensen, 2008). Developmentally, this reasoning process evolves from a concrete operational level (characterized by descriptive categorization of concrete objects and events) to a formal operational level (characterized by generation and testing of causal hypotheses; Lawson, 2004; Lawson, Clark, Cramer-Meldrum, Falconer, Sequist & Kwon, 2000b). Depending on whether or not causal agents are observable, formal reasoning is further conceptualized as consisting of two levels (Lawson et al., 2000b). One involves observable causal agents. For instance, in the case of determining the relationship between the length of a pendulum bob and its period, the causal agent (length) is directly measureable. The other level of formal reasoning involves unseen entities. For example, in the case of testing Mendel's theory, the postulated cause (genes) is not observable, and therefore, a link is needed to connect this unseen agent with experimental outcomes (observable traits). Clearly, reasoning tasks involving

unseen entities are more cognitively demanding than those involving perceptible agents (Lawson, Drake, Johnson, Kwon & Scarpone, 2000c).

In order to successfully carry out a scientific reasoning process, several sub-skills of reasoning are needed. These include proportional reasoning, correlational reasoning, and probabilistic reasoning (Lawson, 2004). While a corpus of previous studies examined scientific reasoning as a single construct, these reasoning sub-skills, particularly their cross-grade progressions from early schoolers to university students, are under-researched.

In a broader context of literature, scientific reasoning is also termed as (or considered as a key component of) scientific thinking, logical thinking, or critical thinking (Facione, 2000; Kuhn, 2002; Zimmerman, 2007). Kuhn (2002), for example, considered scientific thinking as a knowledge-seeking social activity common to the general population. Central to the idea is the coordination between theory and evidence. Note that the definition of theory here is expanded to refer to any mental representations of how things work (Carey, 2000; Gopnik & Wellman, 1994). As Kuhn (2002) noted, four stages of theory-evidence coordination are demonstrated in scientific thinking; they are inquiry, analysis, inference, and argument. At the inquiry stage, activity goals are formulated and questions are asked. Next, at the analysis stage, empirical database is accessed, interpreted, and linked to one's theory (hypothesis). This results in generation and testing of predictions to confirm or disconfirm one's theory at the inference stage. Finally, at the argument stage, evidence-based debates are used to defend one's theory. As seen, this theory-evidence coordination view of scientific thinking intimately relates to the aforementioned scientific reasoning processes in which empirical inquiries are formulated and carried out to test one's postulated causal explanations for an observed phenomenon.

However, the mathematical and logical focus of scientific reasoning by itself cannot represent the entire scientific enterprise. As pointed out by Kuhn (2002), the argument component in theory-evidence coordination highlights the social endeavor of scientific thinking. Similarly, in the new science education framework (NRC, 2012), scientific practices is defined to include not only evidence-based reasoning, such as experimentation and hypothesis testing, but also argumentation and communication (Osborne, 2010; Zimmerman, 2000). In this study, we focus on a central component of scientific practices, that is, scientific reasoning or more specifically, generation and testing of causal hypothesis.

Why Is Scientific Reasoning Important?

The question of why scientific reasoning is important perhaps can be best answered by a review of prior research regarding its effect on student science learning. Vastly different from Inhelder and Piaget (1958), current researchers have shown that scientific reasoning is contextualized and should be considered in concert with one's understanding of the phenomenon under investigation (Kuhn, 2002; Lawson, 2004; Schauble, 1996). Schauble (1996), for instance, conducted a series of interview studies with both elementary schoolers and adults. In her study, each participant was engaged in experimentations over an extended number of sessions to determine the causal mechanisms in two unfamiliar physical science phenomena. She found that how the

participants generated and tested causal explanations was intertwined with their prior knowledge about the situation in question. Specifically, prior knowledge, hypothesis generation, and experimentation strategies interactively influenced each other. In many cases, particularly for the adult participants, increased reasoning strategies near the end of the study due to repeated opportunities increased their domain-related knowledge.

Similarly, a number of studies in various sciences areas confirmed that scientific reasoning, as measured by the CTSR, is a significant predictor for student learning domain concepts. In physics, Coletta and Phillips (2005) administered both CTSR and the Force Concept Inventory (FCI; Hestenes, Wells & Swackhamer, 1992) to students in university-level introductory physics courses. It was found that CTSR strongly correlated with normalized gains on the FCI, a result that was consistently replicated in other studies (Coletta et al., 2007a, b; Moore & Rubbo, 2012; Nieminen, Savinainen & Viiri, 2012). In light of these findings, the researchers claimed that proficient formal reasoners were more likely to achieve higher learning gains than concrete reasoners, and hence, it is important to take student reasoning levels into consideration when assessing their learning gains.

In chemistry, comparable findings were reported. For example, when investigating students' understanding of the particulate nature of matter and its change of state, Tsitsipis and colleagues found that "logical thinking" measured by CTSR can reliably account for a significant portion of the variance in student performance on both conceptual understanding and explanation of the relevant topics (Tsitsipis, Stamovlasis & Papageorgiou, 2010).

In a similar vein of research in biology education, Lawson et al. examined students' performance on CTSR in relation to their performance on biology concept questions (Lawson, Alkhoury, Benford, Clark & Falconer, 2000a). These biology questions target three different types of concepts: descriptive, theoretical, and hypothetical. Descriptive concepts are those that can be derived from direct observations (for example, carnivores). Conversely, theoretical concepts are not derived from observations but are anchored in theories (for example, genes). Finally, hypothetical concepts are at the intermediate level, and although observable in theory under an extended temporal or special frame, they are in reality impossible to be observed (for example, evolution). Lawson et al. divided students into four groups based on their CTSR scores and found that as the reasoning level increased, students were more likely to answer all the three types of questions correctly.

Developmental Framework of Scientific Reasoning

The development of scientific reasoning has been theorized as following a gradual progression through increasingly sophisticated stages. As Lawson (2004) pointed out, this stage-like progression reflects the importance of continuity in the development of scientific reasoning. In other words, reasoning patterns constructed at a later stage utilize (as well as are constrained by) the kinds of reasoning developed in the previous stage. Similarly, Klahr (2000) and Kuhn (2002), after conducting substantive research on scientific thinking among children and adolescents, concluded that coordination of theory and evidence does not emerge abruptly at a single point, but rather it is achieved at successive levels of complexity over an extended period.

The progressive development of scientific reasoning is influenced by multiple factors. Researchers have shown that carefully designed instructional activities that engage students in scientific inquiries—such as designing experiments, generating hypothesis, making inferences, participating in evidence-based arguments, and reflecting on learned ideas at the metacognitive level—can noticeably improve student scientific reasoning skills (Blank, 2000; Fencl, 2010; Gerber, Cavallo & Marek, 2001). Additionally, informal learning experiences, be they related to science or not, can contribute to the advancement of student reasoning (Gerber et al., 2001).

Since scientific reasoning develops continuously and can be potentially enhanced through formal learning, it is imperative that the design and enactment of science instruction closely match the level of student reasoning. To achieve this, a critical, initial step is to understand the progression trend of scientific reasoning across a broader range of grade levels. Drawing on the developmental framework, we conducted a large-scale survey using the CTSR to quantitatively investigate the cross-grade progression trend of Chinese students' scientific reasoning from elementary school to higher education. Results of the study not only reveal the current state of student reasoning at each grade level but can also inform the design and implementation of effective science curricula and instruction at various grades. Additionally, results of our study can be viewed as an initial step toward a better understanding of Chinese science education in the international context and can be of great value for future cross-national comparisons.

Research Questions

Primarily, this study is aimed to investigate the cross-grade trend of student scientific reasoning, as measured by the CTSR. Since CTSR was initially developed through the classical test theory and contains questions targeting various reasoning sub-skills (hypothetical-deductive, proportional, correlational and probabilistic reasoning, and control of variables), a critical issue to be addressed first is the validity and reliability of the CTSR questions, particularly the extent to which these questions can function together to measure the same cognitive construct. Specifically, we seek to answer the following questions: (1) To what extent are the CTSR questions valid and reliable to measure the cognitive construct of scientific reasoning? (2) What is the cross-grade trend of scientific reasoning, as measured by the CTSR, progressing from elementary schoolers to university students? (3) How do the cross-grade progression trends compare among the different reasoning sub-skills measured by the CTSR?

Methods

Student Sample and Settings

The student population of interest in this study is Chinese school attendees from fourth grade up to university senior year. Since the curriculum standards established by the Chinese Department of Education (CDOE) are mandatory, formal instruction to which students are exposed (in terms of both content and pedagogy) is fairly uniform across the nation.

A total number of 2669 students from 13 grade levels participated in the study. The 4th–12th grade students were drawn from nine public schools in both urban and sub-urban areas of Beijing, and the 13th–16th grade students were from two typical universities located in the north and southeast of China. The number of students together with the gender breakdown at each grade is listed in Table 1. Here, grades 4–6 are in primary school, grades 7–9 in middle school, grades 10–12 in high school, and grades 13–16 are of physics major at large research universities in China. Students at the primary level start to take a general science course at grade 4. This course introduces students to basic science ideas to help them make sense of commonly encountered simple phenomena. In middle school, separate science courses aimed at different subject domains begin to emerge. Specifically, biology and geology become independent science courses at grade 7, physics at grade 8, and chemistry at grade 9. These are compulsory courses and continue into high school. Since students at the important milestones of 9th and 12th grades must take local-DOE-administered high school or college entrance examinations (overseen by the CDOE), school instruction is mostly content driven and follows the traditional lecture mode. Similarly, formal education at the tertiary level in China is also delivered in conventional lecture hall environments. To this end, the results of this study by-and-large are a reflection of student scientific reasoning in the context of traditional education.

Instrument

Given its broad use and robust predictive power for student academic performance, the latest version of the CTSR was used in this study [see Coletta and Phillips (2005) for CTSR]. It is a 24-item multiple-choice test designed to measure various scientific reasoning sub-skills, including hypothetical-deductive reasoning, proportional reasoning, correlational reasoning, probabilistic reasoning, and control of variables. Each of the questions on CTSR depicts a science phenomenon and requires the examinee to reason through the situation to predict and explain its outcomes. The English CTSR

Table 1 Sample size and gender composition. Gender information was not collected for grades 13–16

Grade	Sample size	Male (%)	Female (%)	Unidentified (%)
4th	149	49	51	0
5th	349	49	50	1
6th	145	57	43	0
7th	153	52	46	2
8th	330	44	48	8
9th	376	48	51	1
10th	571	45	48	7
11th	230	51	44	5
12th	40	65	35	0
13th	121	–	–	–
14th	69	–	–	–
15th	64	–	–	–
16th	72	–	–	–

was first translated into Mandarin by a group of bilingual science education experts. Two additional bilingual researchers who were not familiar with the CTSR at that time back translated the Mandarin CTSR into English. This back-translated test then was compared against the original CTSR to further refine the Mandarin version. Next, we interviewed over 20 students at different grade levels using the Mandarin CTSR to ensure that the students interpreted the questions as intended. At this stage, we did not detect any major issues that could lead students to misinterpretation of the questions, and the translation of the CTSR was finalized.

The Mandarin CTSR was administered as an in-class paper-and-pencil test to all the participating students within a period of 2 weeks near the middle point of the 2009 fall semester. In each class, students were given 45 min to complete the test and were encouraged to answer the questions seriously according to their true thoughts. No incentives were granted, and students were told that their performance on the CTSR would not affect their course grades.

Rasch Modeling

In order to draw valid inferences from the CTSR results, it is crucial that we establish validity-related evidence for the construct of the test. A Rasch model, which assumes a unidimensionality for a test under investigation, was used to validate the CTSR as well as to estimate student reasoning abilities. Differing from the Classical Test Theory, a Rasch model can yield sample invariant item difficulty and person ability estimates, provided that the sample covers an adequate range of ability and item variations (Bond & Fox, 2007; Boone & Scantlebury, 2005; Liu, 2010). Since all CTSR questions are scored as either correct (1 point) or incorrect (0 point), we use the dichotomous Rasch model for analysis:

$$P(x = 1|\theta_n, \delta_i) = \frac{e^{\theta_n - \delta_i}}{1 + e^{\theta_n - \delta_i}}$$

Here, P is the probability of the nth person correctly answering the ith item, θ is the nth person's ability level, and δ is the ith item's difficulty level. In Rasch modeling, both person ability and item difficulty parameters are estimated. In addition, a set of item fit statistics are generated to reflect the extent to which the individual questions can fit together under the model's unidimensionality assumption. In this study, we examined infit and outfit mean square residuals (MNSQ) for each question. Both fit statistics reveal the divergence between model-estimated results and observed results. The infit MNSQ assigns more weight to situations where person ability is closely matched to item difficulty, whereas the outfit MNSQ assigns equal weight to all cases and therefore is more sensitive to outliers. Typically, infit and outfit MNSQs within the range of [0.7, 1.3] are acceptable (Bond & Fox, 2007). Given the large size of the student sample in the study, we used more stringent upper limits as cutoff points: $\left[0.7, 1 + \frac{2}{\sqrt{N}}\right]$ or [0.7, 1.04] for infit MNSQ and $\left[0.7, 1 + \frac{6}{\sqrt{N}}\right]$ or [0.7, 1.12] for outfit MNSQ, where N is the total number of the student sample (Smith, Schumacker & Bush, 1998). The Winsteps program (Linacre, 2009) was used to establish validity and reliability evidence for the CTSR, and the ConQuest program (Wu, Adams, Wilson & Haldane, 2007) was used to examine the relationships among the reasoning sub-skills and student ability estimates.

Results

Reliability and Construct Validity Evidence of CTSR

To ensure that the results acquired in the study are reliable, we examined two types of reliability derived from the Rasch analysis. One is person separation reliability. As with the traditional Cronbach's alpha, this type of reliability indicates how consistently the CTSR items can separate students into different levels of scientific reasoning. The other type is item separation reliability; it denotes how reliably the test items can be separated into various levels of difficulty. Both types of reliability can vary from 0 to 1, with higher values being desired. In this study, the person and item separation reliability of our dataset are 0.84 and 1.00, respectively, and hence, the subsequent results are considered as reliable.

Rasch analysis also produces estimates of item difficulty and person ability on a common log-odds-unit (logit) scale. Because of this shared scale, a map juxtaposing item and person (also known as a Wright map) can be used to examine how well they match each other. Figure 1 shows a Wright map for the CTSR. As seen, the central axis marked with increasing numbers from bottom to top represents the logit scale. To the left of the scale is a distribution plot of the person estimates. Here, each "#" represents 18 students, and the higher it is, the greater level of scientific reasoning these students have. To the right of the scale is a distribution of the items in terms of their difficulty levels. As with the person distribution, the higher an item is, the more difficult it is. In this study, the difficulty estimates for the CTSR items cover a decent range from −1.82 to +1.42. Although the item distribution appears to be slightly above the person distribution, overall the two sets of estimates match fairly well.

To assess the extent to which the CTSR questions measure the same construct of scientific reasoning and hence to establish evidence of construct validity, we first examined item infit and outfit MNSQs listed in Table 2. Here, the infit MNSQs range from 0.77 to 1.21 and the outfit MNSQs range from 0.71 to 1.61. Except for one question (question 12), all have acceptable infit MNSQs [0.7, 1.04] and outfit MNSQs [0.7, 1.12]. As evident from Fig. 1, question 12 is the most difficulty item on the CTSR. It asks students to explain how they identify and control two variables (light and gravity) to make sense of a given phenomenon regarding flies responding to different experiment settings. Perhaps many students found this question too difficult and therefore guessed on it, causing the fit statistics to fall out of the acceptable range. Despite this finding, the CTSR questions in general can fit the Rasch model reasonably well.

In order to further assess whether or not the items can collectively measure the common construct of science reasoning, we conducted a dimensionality analysis. It was found that the Rasch dimension could explain 62.8% of the total variance in item responses. A principle component analysis of the residuals, after extracting the Rasch dimension, showed that the largest contrast had an eigenvalue of 1.9, equivalent to slightly less than two items of unexplained variance, thus acceptable for assuming a unidimensionality in the CTSR. In other words, no evidence from the analysis suggests that the various reasoning sub-skills (hypothetical-deductive reasoning, proportional reasoning, correlational reasoning, probabilistic reasoning, and control of variables) cannot form a unified cognitive construct. This result provides empirical support for the construct-related validity of the CTSR and buttresses the use of an aggregate score to make inferences about

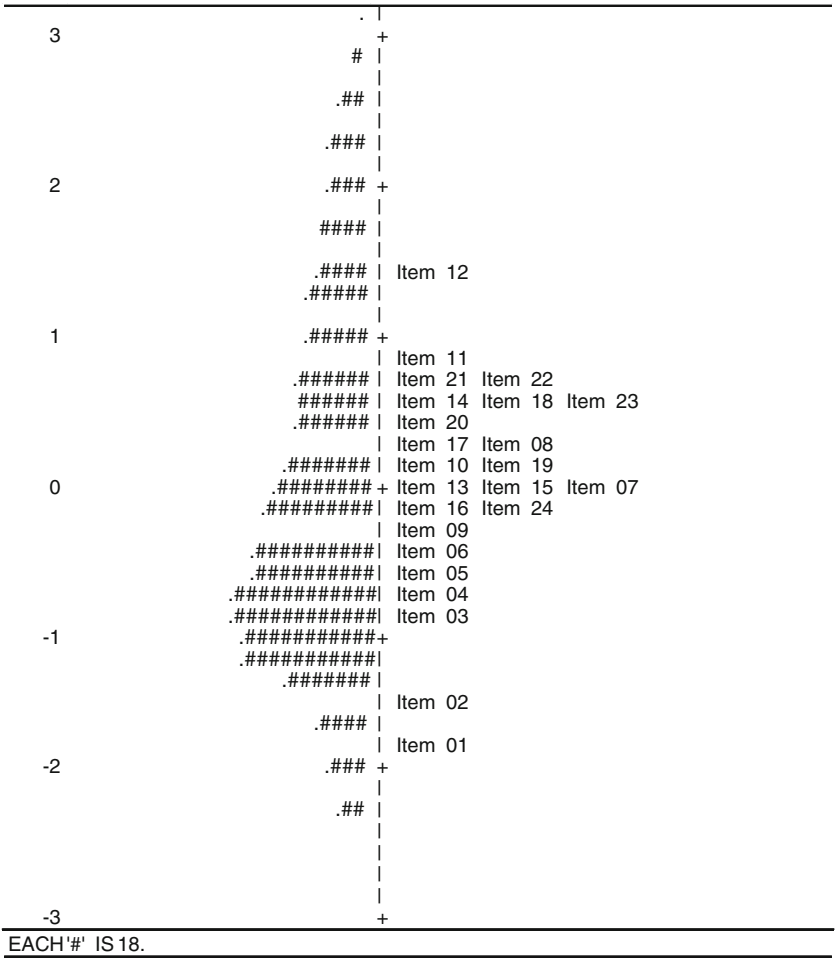

Fig. 1 A Wright map for the CTSR. A vertical logit scale with increasing values from bottom to top separates person ability distribution (*left*) and item difficulty distribution (*right*)

student scientific reasoning measured by the test. It is worth noting that the unidimensionality of CTSR is not at odds with the five sub-skills of reasoning subsumed under the unified construct of the instrument. In fact, the above results indicate that these five sub-skills of reasoning, while each representing an individual sub-scale, can also collectively form a composite and cohesive construct—much in the same way as, for example, the unified construct of "science ability" is composed of four sub-scales of designing and conducting investigations, examining evidence, understanding concepts, and communicating information (Briggs & Wilson, 2003).

Cross-Grade Progression of Scientific Reasoning

Since Rasch analysis generates a person estimate for each student (indicating the individual's scientific reasoning level), we use these estimates to examine the cross-grade progression trend. Average person estimates are computed at each grade level and are plotted

Table 2 Item difficulty and fit statistics of CTSR questions derived from Rasch analysis

Item no.	Measure	Error	MNSQ (infit)	MNSQ (outfit)
1	−1.82	0.05	0.97	0.93
2	−1.56	0.05	0.94	0.91
3	−0.87	0.04	0.83	0.74
4	−0.67	0.04	0.78	0.71
5	−0.55	0.04	0.84	0.77
6	−0.44	0.04	0.77	0.72
7	0.02	0.04	0.79	0.72
8	0.22	0.04	0.97	0.96
9	−0.29	0.04	0.88	0.84
10	0.17	0.04	1.03	1.07
11	0.84	0.05	1.01	1.10
12	1.42	0.05	1.21	1.61
13	−0.05	0.04	1.04	1.05
14	0.54	0.04	0.93	0.90
15	0.04	0.04	0.99	0.97
16	−0.08	0.04	1.01	1.01
17	0.35	0.04	1.01	1.01
18	0.52	0.04	0.96	0.97
19	0.08	0.04	1.02	1.12
20	0.44	0.04	1.03	1.05
21	0.68	0.05	1.04	1.12
22	0.71	0.05	0.98	1.07
23	0.52	0.04	1.03	1.09
24	−0.21	0.04	0.99	0.97
Mean	0.00	0.04	0.96	0.98
SD	0.72	0.00	0.10	0.19

in Fig. 2. Here, the dots represent empirically observed results. Using regression analysis, these observed dots are fitted into various models, such as linear, logistic, logarithmic, and exponential regression models. It is found that the logistic model best fits the observed results, accounting for 89% of the variance in the cross-grade averages ($R^2 = 0.89$).

As seen from Fig. 2, student scientific reasoning measured by the CTSR changes with different rates across the grade levels. At the primary and early secondary levels from grade 4 to grade 7, student reasoning undergo little change despite some fluctuations. Starting from the 8th and 9th grades, an upward progression is noted and this trend continues until the 11th–12th grades. After students enter higher institutions, their scientific reasoning level seems to reach a plateau, remaining more or less stable as they move forward in their education.

Comparison of Progression Trends Among Reasoning Sub-Skills

As mentioned earlier, the CTSR contains questions targeting various scientific reasoning sub-skills, such as hypothetical-deductive reasoning, proportional

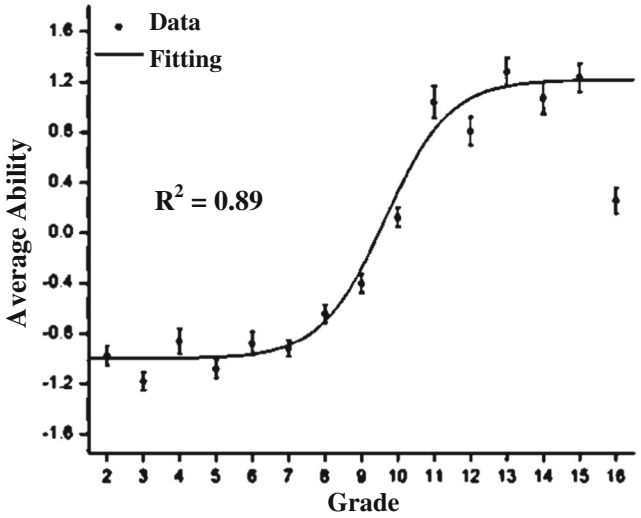

Fig. 2 Cross-grade progression trends of student scientific reasoning measured by the CTSR. *Dots* are empirical averages of Rasch ability at each grade level, *error bars* represent standard errors, and *solid curves* are logistic fitting lines

reasoning, correlational reasoning, probabilistic reasoning, and control of variables. Since these sub-skills represent different levels of complexity (Lawson, 2004; Lawson et al., 2000c), a supposition based on the developmental framework is that they may not progress by following the same trajectory. To test this idea, student performance on each sub-skill is examined.

Student ability estimates on the five reasoning sub-skills were derived from a multidimensional Rasch analysis. For each student, there were five ability estimates corresponding to these sub-skills, respectively. A Wright map displaying the distribution of student ability on each of the sub-skills is depicted in Fig. 3. The person separation indices were 0.79 (proportional reasoning), 0.75 (control of variables), 0.72 (probabilistic reasoning), 0.70 (correlational reasoning), and 0.62 (hypothetical-deductive reasoning), suggesting a decent reliability for a small number of items on each sub-skill.

The average ability estimate for each sub-skill is calculated at each grade level. Figure 4 shows the plots of these results. As before, the dots represent empirical observations, and the curves are best-fit regression lines. Here, the logistic model again accounts for the most portion of the variance in the observed averages for all the five sub-skills ($R^2_{\text{hypothetical-deductive}} = 0.89$, $R^2_{\text{proportional}} = 0.86$, $R^2_{\text{correlational}} = 0.84$, $R^2_{\text{probabilistic}} = 0.82$, $R^2_{\text{control variables}} = 0.87$).

To best capture the differences among various reasoning sub-skills, the regression lines are collectively plotted in Fig. 5. As seen, these cross-grade progression trends share common features as well as exhibit differences. The similarities are primarily manifested at early grade levels. Specifically, prior to grade 8, student performance on all of the five reasoning sub-skills remains low and relatively stable, hence demonstrating little progression. After grade 8, student performance on these sub-skills starts to increase but at different rates. When students enter higher institutions at grade 13, their performance on all the sub-skills seems to reach a plateau, again showing no significant growth. These

	Proportional Reasonig	Controlof Variables	Probabilistic Reasoning	Correlation Reasoning	Hypothetical deductive Reasoning	Items
3						
	X					
	X					
2	XX					
	XXX		X			
	XXX		XX	X		Item 12
	XXXX	X	XX	XX	X	
1	XXXX	X	XXX	XX	XX	Item 11 Item 22
	XXXXXX	XX	XXX	XXX	XXX	Item 14 Item 18 Item 21 Item 23
	XXXXXX	XXXX	X	XXXXX	XXXX	Item 17 Item 20
	XXXXX	XXXXXX	XXXXX	XXXXXX	XXXXXX	Item 08 Item 10 Item 19
0	XXXXX	XXXXXXXXX	XXXXXX	XXXXXXX	XXXXXXXX	Item 07 Item 13 Item 15 Item 16
	XXXX	XXXXXXXXX	XXXXXX	XXXXXX	XXXXXXXX	Item 09 Item 24
	XXX	XXXXXXXXX	XXXXX	XXXXX	XXXXXXXX	Item 06
	XXX	XXXXXXX	XXXX	XXXXX	XXXXXXX	*Item 04* Item 05
-1	XX	XXXX	XXXX	XXXX	XXXX	*Item 03*
	X	XX	XXX	XXXX	XXX	
		X	XXX	XXX	X	*Item 02*
			XX	XXX		
-2			X	XX		*Item 01*
-3						

Each 'X' represents 47.2 cases

Fig. 3 A Wright map of student ability distributions on the five reasoning sub-skills by the CTSR

commonalities are further reflected in the correlation matrix shown in Table 3. As seen, student performances on the five reasoning sub-skills are all positively correlated at a moderate or high level (hence corroborating the previous results regarding the unified construct of scientific reasoning assessed by the CTSR).

However, noticeable differences in these progression trends are evident between grade 8 and grade 12. One major difference lies in the approximate time point at which each trend line starts to grow rapidly. The sub-skill of proportional reasoning is the first that begins to show a noticeable increase at grade 8–9, followed by the sub-skills of probabilistic and correlational reasoning at grade 9–10. However, the sub-skills of control of variable and hypothetical-deductive reasoning seem to have no clear sign of upward climb until grade 10–11, and their progression rates, as seen in the curve slopes, are visibly smaller than those of the other sub-skills. Consequently, near the end of high school, senior year students show dissimilar levels in these five reasoning sub-skills; that is, students perform best on proportional reasoning but worst on control of variable and hypothetical-deductive reasoning.

Discussion

Given the above results, an important question to ask is what are possible causes for these observed patterns? To answer this question, first it is worth noting that the cross-

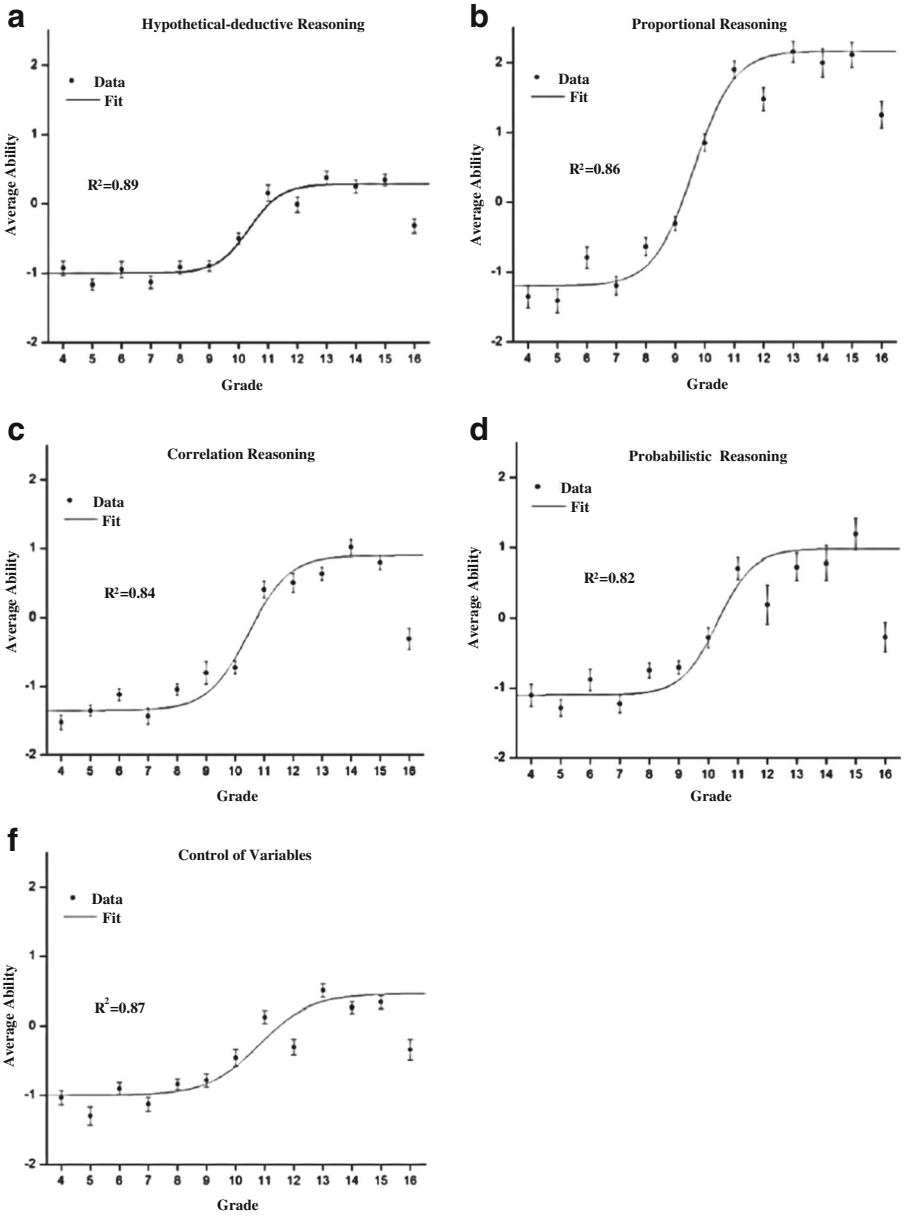

Fig. 4 Cross-grade progression trends of five scientific reasoning sub-skills measured by the CTSR. *Dots* are empirical averages at each grade level, *error bars* represent standard errors, *solid curves* are logistic fitting lines

grade trends emerging from our results conform to the framework regarding the progressive development of scientific reasoning. Although there are fluctuations, the cross-grade progressions we have observed generally follow continuous paths from low to high through increasingly greater levels of improvement.

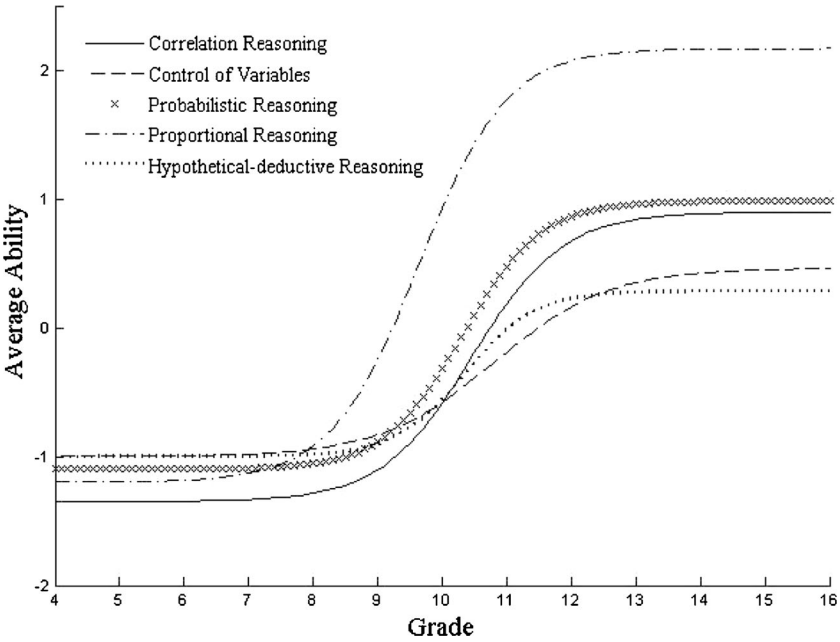

Fig. 5 Logistic-fitted cross-grade progression trends of five scientific reasoning sub-skills on CTSR

For the overall scientific reasoning level measured by the CTSR (Fig. 2), students in the study did not demonstrate a noticeable improvement until the eighth to ninth grades. Multiple factors may have contributed to this increased growth during this period. One is the fact that students at this stage begin to transition from taking one general science course to taking multiple domain-specific science courses, such as biology, geology, physics, and chemistry. These courses, albeit traditionally taught, may have broadened the students' views about the natural world and substantiated their diverse thinking techniques. For instance, in biology and geology courses, students are introduced to microscopic, macroscopic, and large-scale entities, and therefore, they need constantly change perspectives back and forth between these seemingly different worlds to understand otherwise inconceivable phenomena. Similarly, in physics and chemistry classes, students must frequently rely on, for example, proportional thinking to understand relationships between variables or to balance equations.

Table 3 Latent correlation coefficients among the five sub-skills of reasoning measured by the CTSR

Dimension	Proportional	Control of variable	Probabilistic	Correlation	Hypothetical-deductive
Proportional	1				
Control of variable	0.784	1			
Probabilistic	0.614	0.557	1		
Correlation	0.588	0.555	0.616	1	
Hypothetical-deductive	0.606	0.785	0.628	0.56	1

However, formal instruction alone may not solely account for the observed patterns. As noted in the literature, the development of scientific reasoning does not manifest itself as a consequence of a single factor (Klahr, 2000; Kuhn, 2002). Informal experiences, such as extracurricular activities, as well as general development of human intelligence and cognition may also contribute to what we have seen in the above results (Gerber et al., 2001). Nevertheless, this study cannot tease apart these factors.

In terms of the plateau we observed in the period of grades 13–16, it perhaps is due to the fact that students in higher institutions are no longer required to take a range of science courses as broad as at the middle and high school levels. Consequently, the reduced breadth of studies may have slowed down their progress in scientific reasoning. An alternative explanation, however, is that the narrowed focus of their studies may not have a significant impact, but rather the kind of instruction they receive in higher institutions contributes little to the development of their scientific reasoning. This finding also sheds some light on the discordance of science content knowledge with scientific reasoning found among Chinese college students. As revealed from our results, while university students are required to continue learning advanced content knowledge in their specialized fields, their scientific reasoning skills exhibit little improvement (Ding, Wei & Liu, 2016; Ding, Wei & Mollohan, 2016). This perhaps has caused student reasoning to fall out of phase with content learning.

As for the reasoning sub-skills measured by the CTSR, different progression trends are observed. While the sub-skills all start a noticeable growth during the middle school and early high school years, they progress at different paces. It is likely that the emergence of multiple science courses in middle school has facilitated the development of these reasoning sub-skills to different extents. Perhaps, proportional reasoning is frequently invoked in these science classes and therefore displays a fast-paced growth. On the other hand, the sub-skills of control of variables and hypothetical-deductive reasoning are seldom emphasized in the lecture-oriented science classes, and consequently, the development of these skills becomes rather tardy and small. Alternatively, the different cross-grade progression trends also suggest the differing complexity levels of these reasoning sub-skills. For example, learning to control variables or to carry out hypothetical-deductive reasoning can be more challenging than performing proportional reasoning.

Implications for Science Teaching and Learning

Some important information useful for science instruction can be extracted from the above results. First, it is important to acknowledge that what we have observed is a result of multiple factors, and that formal schooling is only part of the story. That said, our results still offer useful suggestions for school instruction. As noted in the previous literature, content knowledge and reasoning are two intertwined, mutually corroborating aspects of science learning (Kuhn, 2002; Osborne, 2010; Schauble, 1996; Zimmerman, 2000). Formal instruction—a dominant means to acquire knowledge and skills in our current society—can inevitably affect the course of how students develop scientific reasoning. In our study, the kind of formal instruction students received is primarily traditional. Therefore, it is conceivable that the cross-grade progression trends we observed are not optimal and have room for improvement.

One area for future improvement derives from the different cross-grade progressions of various reasoning sub-skills. As discussed before, while science instruction at the middle

and high school levels may have contributed to the increased growth of student proportional reasoning, it appears to have created less benefit on the development of hypothetical-deductive reasoning and control of variables. New teaching strategies aimed specifically at these sub-skills are particularly useful. To fulfill this, instructors can purposefully engage students in metacognitive processes to help them identify, articulate, and evaluate their own thoughts related to these reasoning sub-skills. For instance, activities requiring students to explicitly follow the "*if-and-then-but-therefore*" thinking mode can be useful to improve their hypothetical-deductive reasoning (Lawson, 2004).

Notably, as revealed in our study, the middle and high school years seem to be a period of the most rapid growth in student scientific reasoning skills. Therefore, it is crucial that science curriculum and instruction at these grade bands afford students with the best opportunities to advance their reasoning. One possible approach is to incorporate authentic science practices into classroom teaching. In the new framework for K-12 science education (NRC, 2012) and the next generation of science standards (NRC, 2013), scientific and engineering practices are highlighted along with cross-cutting concepts and disciplinary core ideas. Integrating scientific and engineering practices with both disciplinary core ideas and crossing-cutting concepts can maximally improve students' content learning as well as their reasoning skills.

Another possible area for improvement is formal instruction in higher education. Note that the plateau in the trend lines we observed is not due to a ceiling effect; in other words, there is room for students at this level to improve scientific reasoning. Given the fact that tertiary-level educators in general have more freedom in selecting content and pedagogies, more hands-on and minds-on activities can be introduced into classroom to afford students with opportunities of authentic scientific practices. Additionally, offering forums that allow students to critically evaluate science-related media reports through evidence-based argumentation may be profitable.

Finally, science instruction at the primary level is also an area for improvement. In our study, students before grade 8 made no significant cross-grade progression on all the reasoning sub-skills. Although students at this level still struggle with abstract thinking, previous studies have found that with sufficient practice, primary schoolers are able to perform reasoning about complex situations (Zimmerman, 2000). To this end, frequently engaging students in reasoning that involves multiple variables may be useful. Depending on the grade levels, an instructor can also provide suitable scaffolding to assist students in carrying out these tasks.

Limitations of Current Work and Implications for Future Research

As with any empirical work, this study contains several limitations. What is worth noting first perhaps is the cross-sectional nature of the study. Instead of performing a repeated measurement with the same group of students over an extended period of time, we took a snap shot of different student groups across a wide range of grades at approximately the same time. Results using the former approach (longitudinal study) may differ from those in the latter (cross-sectional study). However, several factors inherent in our study can mitigate this concern. First, the mandatory school curriculum standards in China create a relative uniformity in formal education across the nation. Additionally, abrupt large-scale changes in either formal education or student population are unlikely to occur in China within a short time frame. Hence, what we see today

in lower-grade students (in terms of their backgrounds and learning environments) is more-or-less equivalent to what higher-grade students experienced several years ago. In other words, there is a sense of uniformity and stability for the context of the study. This in fact is further reinforced by our surveying a large sample of students and averaging the results at each grade level for analysis. As shown by Bates, Galloway, Loptson and Slaughter (2011) who investigated students' epistemology about physics and learning, longitudinal and cross-sectional studies yielded similar results. What is more, the fact that the cross-grade progressions are examined by using Rasch estimates known to be less sample-dependent adds further stability to our results. That said, divergences to some degree may still exist between the two approaches. Future research can verify and explicate such possibilities.

Another limitation of the study is the uneven distribution of participant numbers across the grades. In particular, there were less than a hundred students in the 12th, 14th, 15th, and 16th grades. This issue is primarily due to logistic constraints. As mentioned earlier, students at grade 12 must take a college entrance examination (one that is considered by many students and their parents as the most important examination) before continuing higher education. Teachers and school administrators therefore are highly protective of their classroom; research activities—however relevant to content learning—are often deemed as a distraction and interruption. Consequently, we only managed to recruit 40 students at this grade level. As for the university students, the number of participants was constrained by the available undergraduates in a major. Here, we only surveyed physics majors at two large research universities in China. Future work can seek to expand the study to other majors and institutions.

Also worth noting is the fact that this study focuses on the collective progression of students' scientific reasoning from lower to higher grades. What has not been sought is the progression of individual students and its mechanism. To address this issue, smaller-scale qualitative studies such as case studies may be a useful future direction.

Conclusions

We conducted a large-scale survey of Chinese students' scientific reasoning using the CTSR across 13 grade levels, from elementary school to university. The construct-related validity of the CTSR was first evaluated through Rasch analysis. Results show that the CTSR measures are valid and that various scientific reasoning sub-skills measured by the CTSR questions in general can form a cohesive cognitive construct. Additionally, results derived from the CTSR were also shown to have a high reliability. This allowed us to proceed using the CTSR to examine cross-grade progressions in student reasoning skills. It was found that the cross-grade progressions follow a continuous path with increasing improvement. In particular, all of the reasoning sub-skills begin to show a significant growth during the middle and high school years, but at different rates. The development of proportional reasoning is relatively fast, but the progression of hypothetical-deductive reasoning and control of variables is slow-paced and small. Multiple factors may have contributed to the observed results. Given the traditional nature of formal school instruction in China, the results represent a less optimal cross-grade progression in student scientific reasoning. Future work in science education at various levels can be conducted to improve the status quo.

Acknowledgement This work is partially supported by the National Science Foundation (Grant No. DRL - 1252399) and The Ohio State University ASC/EHE Seed Grant (2016–2017).

References

Ates, S. & Cataloglu, E. (2007). The effects of students' reasoning abilities on conceptual understandings and problem-solving skills in introductory mechanics. *European Journal of Physics*, 28(6), 1161–1171.

Bao, L., Cai, T., Koenig, K.M., Fang, K., Han, J., Wang, J. et al. (2009). Learning and scientific reasoning. *Science*, 323(5914), 586–587.

Bates, S., Galloway, R., Loptson, C. & Slaughter, K. (2011). How attitudes and beliefs about physics change from high school to faculty. *Physical Review Special Topics - Physics Education Research*, 7(020114), 1–8.

Blank, L.M. (2000). A metacognitive learning cycle: A better warranty for student understanding? *Science Education*, 84(4), 486–506.

Bond, T.G. & Fox, C.M. (2007). Applying the Rasch model: Fundamental measurement in the human sciences (2nd ed.). New York, NY: Routledge, Taylor & Francis.

Boone, W.J. & Scantlebury, K. (2005). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education*, 90(2), 253–269.

BouJaoude, S., Salloum, S. & Abd-El-Khalick. (2004). Relationships between selective cognitive variables and students' ability to solve chemistry problems. *International Journal of Science Education*, 26(1), 63–84.

Briggs, D. & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement*, 4(1), 87–100.

Carey, S. (2000). Science education as conceptual change. *Journal of Applied Developmental Psychology*, 21(1), 13–19.

Coletta, V. P. & Phillips, J. A. (2005). Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability. *American Journal of Physics*, 73(12), 1172–1182.

Coletta, V.P., Phillips, J.A. & Steinert, J.J. (2007a). Interpreting force concept inventory scores: Normalized gain and SAT scores. *Physical Review Special Topics - Physics Education Research*, 3(1), 010106.

Coletta, V.P., Phillips, J.A. & Steinert, J.J. (2007b). Why you should measure your students' reasoning ability. *Physics Teacher*, 45(4), 235–238.

Ding, L. (2014). Verification of causal influences of reasoning skills and epistemology on physics conceptual learning. *Physical Review Special Topics - Physics Education Research*, 10(2), 023101. doi:10.1103/PhysRevSTPER.10.023101.

Ding, L., Wei, X. & Liu, X. (2016). Variations in university students' scientific reasoning skills across majors, years, and types of institutions. *Research in Science Education*, 46(5), 613-632. doi:10.1007/s11165-015-9473-y.

Ding, L., Wei, X. & Mollohan, K. (2016). Does higher education improve student scientific reasoning skills? *International Journal of Science and Mathematics Education*, 14(4), 619–634. doi:10.1007/s10763-014-9597-y.

Facione, P.A. (2000). The disposition toward critical thinking: Its character, measurement, and relationship to critical thinking skills. *Informal Logic*, 20(1), 61–84.

Fencl, H.S. (2010). Development of students' critical-reasoning skills through content-focused activities in a general education course. *Journal of College Science Teaching*, 39(5), 56–62.

Gerber, B.L., Cavallo, A.M.L. & Marek, E.A. (2001). Relationships among informal learning environments, teaching procedures and scientific reasoning ability. *International Journal of Science Education*, 23(5), 535–549.

Gopnik, A. & Wellman, H.M. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind* (pp. 257–293). New York, NY: Cambridge University.

Hestenes, D., Wells, M. & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, 30(3), 141–158.

Inhelder, B. & Piaget, J. (1958). The growth of logical thinking from childhood to adolescence. New York, NY: Basic.

Klahr, D. (2000). Exploring science: The cognitive and development of discovery processes. Cambridge, MA: MIT Press.

Kuhn, D. (2002). What is scientific thinking and how does it develop? In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 371–393). Malden, MA: Blackwell Publishers.
Kuhn, D., Black, J., Keselman, A., & Kaplan, D. (2000). The development of cognitive skills to support inquiry learning. *Cognition and Instruction*, 18(4), 495–523.
Lawson, A.E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching*, 15(1), 11–24.
Lawson, A.E. (2000). The generality of hypothetico-deductive reasoning: Making scientific thinking explicit. *American Biology Teacher*, 62(7), 482–495.
Lawson, A.E. (2004). The nature and development of scientific reasoning. *International Journal of Science and Mathematics Education*, 2(3), 307–338.
Lawson, A.E. (2005). What is the role of induction and deduction in reasoning and scientific inquiry? *Journal of Research in Science Teaching*, 42(6), 716–740.
Lawson, A.E., Alkhoury, S., Benford, R., Clark, B.R., & Falconer, K.A. (2000a). What kinds of scientific concepts exist? Concept construction and intellectual development in college biology. *Journal of Research in Science Teaching*, 37(9), 996–1018.
Lawson, A.E., Clark, B., Cramer-Meldrum, E., Falconer, K.A., Sequist, J.M. & Kwon, Y.-J. (2000b). Development of scientific reasoning in college biology: Do two levels of general hypothesis-testing skills exist? *Journal of Research in Science Teaching*, 37(1), 81–101.
Lawson, A.E., Drake, N., Johnson, J., Kwon, Y.-J. & Scarpone, C. (2000c). How good are students at testing alternative explanations of unseen entities? *American Biology Teacher*, 62(4), 249–255.
Lawson, A.E., Oehrtman, M. & Jensen, J. (2008). Connecting science and mathematics: The nature of scientific and statistical hypothesis testing. *International Journal of Science and Mathematics Education*, 6(2), 405–416.
Linacre, J.M. (2009). A user's guide to Winsteps, Rasch measurement program. Chicago, IL: MESA Press.
Liu, X. (2010). Using and developing measurement instruments in science education: A Rasch modeling approach. Charlotte, NC: Information Age Publishing.
Meltzer, D.E. (2002). The relationship between mathematics preparation and conceptual learning gains in physics: A possible "hidden variable" in diagnostic pretest scores. *American Journal of Physics*, 70(12), 1259–1268.
Moore, J.C. & Rubbo, L.J. (2012). Scientific reasoning abilities of nonscience majors in physics-based courses. *Physical Review Special Topics-Physics Education Research*, 8(1), 010106.
Nieminen, P., Savinainen, A. & Viiri, J. (2012). Relations between representational consistency, conceptual understanding of the force concept, and scientific reasoning. *Physical Review Special Topics - Physics Education Research*, 8(010123), 1–10.
National Research Council (2000). Inquiry and the national science education standards: A guide for teaching and learning. Washington, DC: The National Academies Press. doi:10.17226/9596.
National Research Council (2012). A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. Washington, DC: The National Academies Press. doi:10.17226/13165.
National Research Council (2013). Next generation science standards: For States, by States. Washington, DC: The National Academies Press. doi:10.17226/18290.
Osborne, J. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *Science*, 328(23), 463–466.
Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology*, 32(1), 102–119.
Shayer, M. & Adey, P.S. (1993). Accelerating the development of formal thinking in middle and high school students IV: Three years after a two-year intervention. *Journal of Research in Science Teaching*, 30(4), 351–366.
Smith, R., Schumacker, R. & Bush, M.J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, 2(1), 66–78.
Tsitsipis, G., Stamovlasis, D. & Papageorgiou, G. (2010). The effect of three cognitive variables on students' understanding of the particulate nature of matter and its changes of state. *International Journal of Science Education*, 32(8), 987–1016.
Wu, M.L., Adams, R.J., Wilson, M.R. & Haldane, S.A. (2007). ACER ConQuest version 2.0: Generalized item response modeling software. Camberwell, Victoria: Australian Council for Educational Research Ltd.
Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review*, 20(1), 99–149.
Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, 27(2), 172–223.