PHYSICS

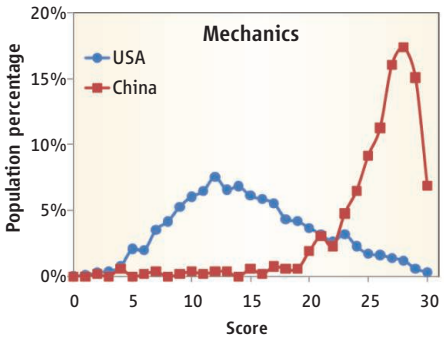
Learning and Scientific Reasoning

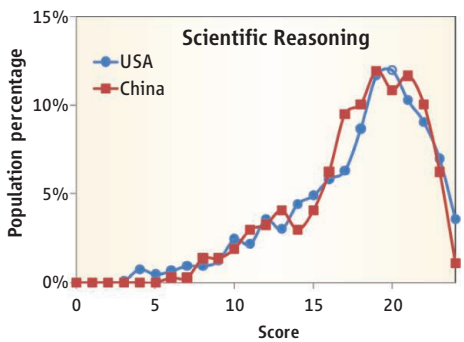
Comparisons of Chinese and U.S. students show that content knowledge and reasoning skills diverge.

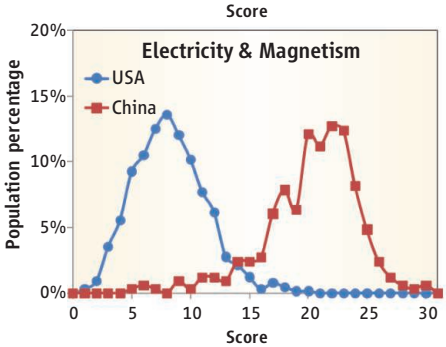
Lei Bao,[1]* Tianfan Cai,[2] Kathy Koenig,[3] Kai Fang,[4] Jing Han,[1] Jing Wang,[1] Qing Liu,[1] Lin Ding,[1] Lili Cui,[5] Ying Luo,[6] Yufeng Wang,[2] Lieming Li,[7] Nianle Wu[7]

The development of general scientific abilities is critical to enable students of science, technology, engineering, and mathematics (STEM) to successfully handle open-ended real-world tasks in future careers (*1*–*6*). Teaching goals in STEM education include fostering content knowledge and developing general scientific abilities. One such ability, scientific reasoning (*7*–*9*), is related to cognitive abilities such as critical thinking and reasoning (*10*–*14*). Scientific-reasoning skills can be developed through training and can be transferred (*7*, *13*). Training in scientific reasoning may also have a long-term impact on student academic achievement (*7*). The STEM education community considers that transferable general abilities are at least as important for students to learn as is the STEM content knowledge (*1*–*4*). Parents consider science and mathematics to be important in developing reasoning skills (*15*).

We therefore asked whether learning STEM content knowledge does in fact have an impact on the development of scientific-reasoning ability. The scientific-reasoning ability studied in this paper focuses on domain-general reasoning skills such as the abilities to systematically explore a problem, to formulate and test hypotheses, to manipulate and isolate variables, and to observe and evaluate the consequences.







TEST SCORES (%)			
Test	*China (n)*	*USA (n)*	*Effect size*
FCI	85.9 ± 13.9 (523)	49.3 ± 19.3 (2681)	1.98
BEMA	65.6 ± 12.8 (331)	26.6 ± 10.0 (650)	3.53
LCTSR	74.7 ± 15.8 (370)	74.2 ± 18.0 (1061)	0.03

Content knowledge and reasoning skills diverge. Comparisons of U.S. and Chinese freshmen college students show differences on tests of physics content knowledge but not on tests of scientific reasoning.

Research Design

Students in China and the United States go through very different curricula in science and mathematics during their kindergarten through 12th grade (K–12) school years. This provides systemically controlled long-term variation on STEM content learning, which we used to study whether or not such learning has any impact on the development of scientific-reasoning ability. Scientific reasoning is not explicitly taught in schools in either country.

In China, K–12 education is dominated by the nationwide college admission exam given at the end of grade 12. To comply with the requirements of this exam, all Chinese schools adhere to a national standard within all courses. In physics, for example, every student goes through the same physics courses, which start in grade 8 and continue every semester through grade 12, providing 5 years of continuous training on introductory physics topics (*16*). The courses are algebra-based with emphasis on development of conceptual understanding and skills needed to solve problems.

In contrast, K–12 physics education in the United States is more varied. Although students study physics-related topics within other general science courses, only one of three high school students enrolls in a two-semester physics course (*17*). As a result, the amount of instructional time and the amount of emphasis on conceptual physics understanding and problem-solving skills are very different in the two countries. Similar curriculum differences between the United States and China are reflected in other STEM areas such as chemistry, biology, and mathematics (*16*).

Chinese students go through rigorous problem-solving instruction in all STEM subject areas throughout most of their K–12 school years and become skillful at solving content-based problems. It remains unclear, however, whether this training is transferable beyond the specific content areas and problem types taught.

We used quantitative assessment instruments (described below) to compare U.S. and Chinese students' conceptual understanding in physics and general scientific-reasoning ability. Physics content was chosen because the subject is conceptually and logically sophisticated and is commonly emphasized in science education (*15*). Assessment data were collected from both Chinese and U.S. freshmen college students before college-level physics instruction. In this way the data reflect students' knowledge

[1]Department of Physics, The Ohio State University, Columbus, OH 43210, USA. [2]Department of Physics, Beijing Jiaotong University, Beijing 100044, China. [3]Department of Physics, Wright State University, Dayton, OH 45435, USA. [4]Department of Physics, Tongji University, Shanghai 200092, China. [5]Department of Physics, University of Maryland, Baltimore County, Baltimore, MD 21250, USA. [6]Department of Physics, Beijing Normal University, Beijing 100875, China. [7]Department of Physics, Tsinghua University, Beijing 100084, China.

*Author for correspondence. E-mail: bao.15@osu.edu

Published by AAAS

and skill development from their formal and informal K–12 education experiences.

Data Collection and Analysis

From the early 1980s, researchers and educators in psychology and cognitive science (*11*–*14*) have developed many quantitative instruments that assess reasoning ability. Some are included as components in standard assessments such as the Graduate Record Examination, whereas others are stand-alone tests such as Lawson's *Classroom Test of Scientific Reasoning* (LCTSR) (*8*, *9*). We used the LCTSR because of its popularity among STEM educators and researchers. Common categories of reasoning ability assessments include proportional reasoning, deductive and inductive reasoning, control of variables, probability reasoning, correlation reasoning, and hypothesis evaluation, all of which are crucial skills needed for a successful career in STEM.

Research-based standardized tests that assess student STEM content knowledge are also widespread. For example, in physics, education research has produced many instruments. We used the *Force Concept Inventory* (FCI) (*18*, *19*) and the *Brief Electricity and Magnetism Assessment* (BEMA) (*20*). These tools are regularly administered by physics education researchers and educators to evaluate student learning of specific physics concepts.

Using FCI (mechanics), BEMA (electricity and magnetism), and LCTSR (scientific reasoning), we collected data (see figure, page 586) from students (N = 5760) in four U.S. and three Chinese universities. All the universities were chosen to be of medium ranking (*15*). The students tested were freshmen science and engineering majors enrolled in calculus-based introductory physics courses. The tests were administered before any college-level instruction was provided on the related content topics. The students in China used Chinese versions of the tests, which were first piloted with a small group of undergraduate and graduate students (n = 22) to remove language issues.

The FCI results show that the U.S. students have a broad distribution in the medium score range (from 25 to 75%). This appears to be consistent with the educational system in the United States, which produces students with a blend of diverse experiences in physics learning. In contrast, the Chinese students had all completed an almost identical extensive physics curriculum spanning five complete years from grade 8 through grade 12. This type of education background produced a narrow distribution that peaks near the 90% score.

For the BEMA test, the U.S. students have a narrow distribution centered a bit above the chance level (chance 20%). The Chinese students also scored lower than their performance on the FCI, with the score distribution centered around 70%. The lower BEMA score of students in both countries is likely due to the fact that some of the topics on the BEMA test (for example, Gauss's law) are not included in standard high school curricula.

The FCI and BEMA results suggest that numerous and rigorous physics courses in the middle and high school years directly affect student learning of physics content knowledge and raise students to a fairly high performance level on these physics tests.

The results of the LCTSR test show a completely different pattern. The distributions of the Chinese and U.S. students are nearly identical. Analyses (*15*) suggest that the similarities are real and not an artifact of a possible ceiling effect. The results suggest that the large differences in K–12 STEM education between the United States and China do not cause much variation in students' scientific-reasoning abilities. The results from this study are consistent with existing research, which suggests that current education and assessment in the STEM disciplines often emphasize factual recall over deep understanding of science reasoning (*2*, *21*–*23*).

What can researchers and educators do to help students develop scientific-reasoning ability? Relations between instructional methods and the development of scientific reasoning have been widely studied and have shown that inquiry-based science instruction promotes scientific-reasoning abilities (*24*–*29*). The current style of content-rich STEM education, even when carried out at a rigorous level, has little impact on the development of students' scientific-reasoning abilities. It seems that it is not what we teach, but rather how we teach, that makes a difference in student learning of higher- order abilities in science reasoning. Because students ideally need to develop both content knowledge and transferable reasoning skills, researchers and educators must invest more in the development of a balanced method of education, such as incorporating more inquiry-based learning that targets both goals.

Our results also suggest a different interpretation of assessment results. As much as we are concerned about the weak performance of American students in TIMSS and PISA (*30*, *31*), it is valuable to inspect the assessment outcome from multiple perspectives. With measurements on not only content knowledge but also other factors, one can obtain a more holistic evaluation of students, who are indeed complex individuals.

References and Notes

1. R. Iyengar *et al*., *Science* **319**, 1189 (2008).
2. A. Y. Zheng *et al*., *Science* **319**, 414 (2008).
3. B. S. Bloom, Ed., *Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook I: Cognitive Domain* (David McKay, New York 1956).
4. National Research Council (NRC), *National Science Education Standards* (National Academies Press, Washington, DC, 1996).
5. NRC, *Learning and Understanding: Improving Advanced Study of Mathematics and Science in U.S. High Schools* (National Academies Press, Washington, DC, 2002).
6. H. Singer, M. L. Hilton, H. A. Schweingruber, Eds., *America's Lab Report* (National Academies Press, Washington, DC, 2005).
7. P. Adey, M Shayer, *Really Raising Standards: Cognitive Intervention and Academic Achievement* (Routledge, London, 1994).
8. A. E. Lawson, *J. Res. Sci. Teach*. **15**, 11 (1978).
9. Test used in this study was *Classroom Test of Scientific Reasoning*, rev. ed. (2000).
10. P. A. Facione, *Using the California Critical Thinking Skills Test in Research, Evaluation, and Assessment* (California Academic Press. Millbrae, CA, 1991).
11. H. A. Simon, C. A. Kaplan in *Foundations of Cognitive Science*s, M. I. Posner, Ed. (MIT Press, Cambridge, MA, 1989), pp. 1–47.
12. R. E. Nisbett, G. T. Fong, D. R. Lehman, P. W. Cheng, *Science* **238**, 625 (1987).
13. Z. Chen, D. Klahr, *Child Dev.* **70**, 1098 (1999).
14. D. Kuhn, D. Dean, *J. Cognit. Dev.* **5**, 261 (2004).
15. See Supporting Online Material for more details.
16. This is based on the Chinese national standards on K–12 education (www.pep.com.cn/cbfx/cpml/).
17. J. Hehn, M. Neuschatz, *Phys. Today* **59**, 37 (2006).
18. D. Hestenes, M. Wells, G. Swackhamer, *Phys. Teach.* **30**, 141 (1992).
19. The test used in this study is the 1995 version.
20. L. Ding, R.Chabay, B. Sherwood, R. Beichner, *Phys. Rev. ST Phys. Educ. Res.* **2**, 010105 (2006).
21. M. C. Linn *et al*., *Science* **313**, 1049 (2006).
22. A. Schoenfeld, *Educ. Psychol.* **23**, 145 (1988).
23. A. Elby, *Am. J. Phys*. **67**, S52 (1999).
24. C. Zimmerman, *Dev. Rev.* **27**, 172 (2007).
25. P. Adey, M. Shayer, *J. Res. Sci. Teach.* **27**, 267 (1990).
26. A. E. Lawson, *Science Teaching and the Development of Thinking* (Wadsworth, Belmont, CA, 1995).
27. R. Benford, A. E. Lawson, *Relationships Between Effective Inquiry Use and the Development of Scientific Reasoning Skills in College Biology Labs* (Arizona State University, Tempe, AZ, 2001); Educational Resources Information Center (ERIC) accession no. ED456157.
28. E. A. Marek, A. M. L. Cavallo, *The Learning Cycle and Elementary School Science* (Heinemann, Portsmouth, NH, 1997).
29. B. L. Gerber, A. M. Cavallo, E. A. Marek, *Int. J. Sci. Educ.* **23**, 5359 (2001).
30. Trends in International Mathematics and Science Study (TIMSS), http://nces.ed.gov/timss/.
31. Programme for International Student Assessment (PISA), www.pisa.oecd.org/.

32. We wish to thank all the teachers who helped with this research.

Supporting Online Material

www.sciencemag.org/cgi/content/full/323/5914/586/DC1

10.1126/science.1167740

Originally posted 29 January 2009, corrected 4 February 09

www.sciencemag.org/cgi/content/full/323/5914/586/DC1

Supporting Online Material for

Learning and Scientific Reasoning

Lei Bao,* Tianfan Cai, Kathy Koenig, Kai Fang, Jing Han, Jing Wang, Qing Liu, Lin Ding, Lili Cui, Ying Luo, Yufeng Wang, Lieming Li, Nianle Wu

*To whom correspondence should be addressed. E-mail: bao.15@osu.edu

Published 30 January 2009, *Science* **323**, 586 (2009)
DOI: 10.1126/science.1167740

This PDF file includes

Correction 4 February 2009: A section on defining scientific reasoning with references was added to the revision that was submitted before the posting date.

Learning and Scientific Reasoning
Supporting Online Materials

Definition of Scientific Reasoning

In the literature, there are many definitions of scientific reasoning. From the science literacy perspective (*S1*, *S2*), scientific reasoning represents the cognitive skills necessary to understand and evaluate scientific information, which often involve understanding and evaluating theoretical, statistical, and causal hypotheses.

From the research point of view (*S3*), scientific reasoning, broadly defined, includes the thinking and reasoning skills involved in inquiry, experimentation, evidence evaluation, inference, and argumentation that support the formation and modification of concepts and theories about the natural and social world. Two main types of knowledge, namely, domain-specific knowledge and domain-general strategies, have been widely researched (*S3*).

Specifically, the measurement instrument used in this paper, the Lawson's Classroom Test of Scientific Reasoning (*S4*), assesses students' abilities in six dimensions including conservation of matter and volume, proportional reasoning, control of variables, probability reasoning, correlation reasoning, and hypothetical-deductive reasoning. These skills are important concrete components of the broadly defined scientific reasoning ability (*S5*-*S9*); therefore, in this paper scientific reasoning is operationally defined in terms of students' ability in handling questions of the six skill dimensions.

Views and Expectations on How to Improve Scientific Reasoning

The view of teachers and the general public on what helps the development of scientific reasoning is an important issue, since it may influence, either explicitly or implicitly, how we educate our next generation.

Through informal discussions with people of a wide variety of backgrounds including teachers, undergraduate and graduate students, scientists, and people from the general public ($N_{\text{Total}} \approx 50$), we have observed that most of them believed that more science and mathematics courses will improve students' scientific reasoning abilities. To obtain a quantitative measure of the popularity of this belief, we developed a survey on people's views concerning science learning and scientific reasoning. We include pilot data here to provide an empirical baseline result on one of the survey questions that directly addresses the question of interest.

The Survey Question:

How much do you think learning science and mathematics in schools will play a role in developing students' reasoning ability? (Circle one below)

A. ***About 100%*** *(the development of students' reasoning ability benefits entirely from learning science and mathematics in schools)*
B. ***About 80%*** *(the development of students' reasoning ability benefits mostly from learning science and mathematics in schools)*

C. ***About 50%*** *(the development of students' reasoning ability benefits from learning science and mathematics in schools and other activities, both of which are about equally important)*

D. ***About 20%*** *(the development of students' reasoning ability benefits only slightly from learning science and mathematics in schools)*

E. ***About 0%*** *(the development of students' reasoning ability doesn't benefit from learning science and mathematics in schools at all)*

This question was given to pre-service teachers (sophomore college students) in both U.S.A. and China. Students' responses are summarized in Table S1.

Table S1. Survey results on views about science learning and scientific reasoning.

Answers	Science and math's effect on reasoning ability (%)	
	USA (n = 25)	China (n = 28)
A	15	0
B	54	82
C	31	18
D	0	0
E	0	0
Weighted sum of impact*	74	75

* The weighted sum of impact is computed as the sum of the products of the population percentage of the answers and the impact values specified in the answers.

The results suggest that although the distributions of answers are different, both populations have a similar overall rating regarding the role that learning science and mathematics plays in developing students' reasoning abilities.

Possible Ceiling Effect of the Lawson Test:

The Lawson test measures fundamental reasoning components with simple context scenarios that do not require complex content understanding. This test design can improve the measurement of the basic reasoning abilities by reducing the possible interference from understandings of content knowledge. The test results of college students, which average around 75% on the test, indicate a possible ceiling effect. To understand the impact of the ceiling effect in this study, we conducted further research to measure how the scientific reasoning ability is developed through the school and college years. We collected data with Chinese students from 3rd grade to second-year college level (N_{Total} = 6258). The students are from 141 classes in 20 schools from eight regions around China; thus, they form a more representative population. The results are plotted in Figure S1. The red dots are grade-average LCTSR scores (out of 24). The red line is referred as a "Learning Evolution Index Curve (LEI-Curve)", which is obtained by fitting the data with a logistic function motivated by item response theory (*S10*):

$$y = F + \frac{C - F}{1 + e^{-\alpha(x-b)}} \quad (1)$$

where

x – Student grade level
y – Student score
F – Floor, the lowest score score possible
C – Ceiling, the highest score possible
α – Discrimination factor, which controls the steepness of the curve.
b – Grade-based item difficulty, which controls the center of the curve.

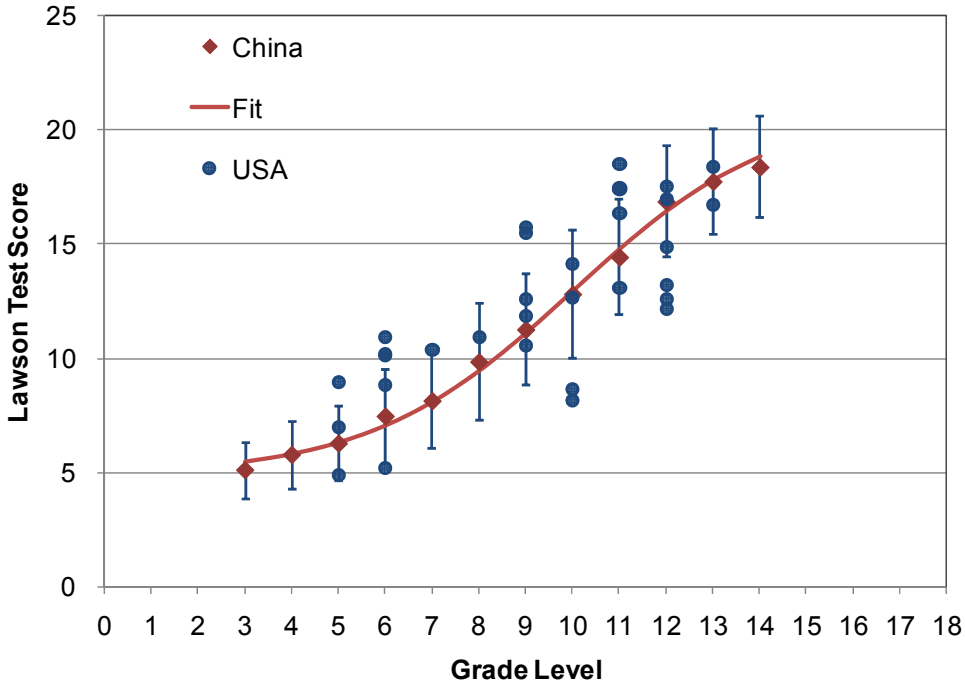


Figure S1. The developmental trend of Chinese and U.S. students' LCTSR scores (out of 24).

The results shown here are for the purpose of presenting the general developmental trend of the Lawson test scores of Chinese and U.S. students. The error bars shown in the graph are standard deviations of the class mean scores, which gives the range of variance if one were to compare the mean scores of different classes.

The U.S. data were collected in a Midwestern state from 30 classes of students across 14 private and public schools (N_{Total} = 1078). Each blue dot on the graph represents a group of students within the same science course, such as a biology or a chemistry course. Owing to the relatively small sample size, we plotted the class group mean scores of the U.S. data in blue dots on top of the Chinese data. We can see that from 5$^{\text{th}}$ grade to first-year college level, the U.S. and Chinese data are within one standard deviation of each other, showing a similar developmental scale.

To obtain a more quantitative measure, we combined the students in 11$^{\text{th}}$ and 12$^{\text{th}}$ grades and computed the average LCTSR scores for both U.S and Chinese populations. The results are

summarized in Table S2, which show little difference between students from the two countries. Since the average scores are at 63% level, the results are less affected by the possible ceiling effect. Therefore, based on the data from both college students and high school students, we can conclude that the similar performance of U.S. and Chinese students on the LCTSR represents a real signal rather than an artifact of the ceiling effect.

Although the reasoning abilities tested in the LCTSR appear to be simple to expert scientists, these are crucial fundamental components for more sophisticated skills. For example, the ability to "control variables" is involved in scientific experimentation and modeling at all levels and has been widely studied (*S5*, *S8*). Being able to fully develop this skill is a crucial step for developing more advanced higher-order abilities. The developmental data show that students start to fully develop the basic reasoning abilities around their college years. However, in order to assess the reasoning abilities of senior college students and graduate students, we need to develop questions that involve more advanced reasoning components.

Table S2. LCTSR mean scores of 11th and 12th graders combined. Average ± SD.

Populations	No. of classes	Students N_{Total}	Class means (%)	Population (%)
USA	11	402	64.2 ± 9.7	64.2 ± 16.3
China	39	1786	62.6 ± 10.0	62.7 ± 15.6

Sample Backgrounds and Further Analysis

The comparision groups are freshmen college students of science and engineering majors enrolled in entry level calculus-based physics courses. These groups of students form the main body of the next generation technology workforce in both U.S.A. and China.

Data from four U.S. universities and three Chinese universities are used in the paper. The four U.S. universities are labled U1, U2, U3, and U4. University ranking and backgrounds are given below (based on 2007 U.S. News and World Report Ranking):

- U1 is a large research-1 state university, U.S. ranking top 60, acceptance rate 59%.
- U2 is a large research-1 state university, U.S. ranking top 60, acceptance rate 60%.
- U3 is a large tier-4 state university with an acceptance rate of 84%.
- U4 is a large tier-3 state university with an acceptance rate of 69%.

The three Chinese universities are labeled with C1, C2, and C3. Their national rankings are also given below (based on 2007 ranking from Sina Educatoin News, http://edu.sina.com.cn). (A national university is one that is under direct control of the department of education.)

- C1 is a top 30 national university.
- C2 is a top 60 national university.
- C3 is a top 130 national university.

In the selection of universities, we targeted the ones with medium ranking in order to make a more representative pool of the population. The summary of data from all samples is in Tables S3 and S4. The data were mostly collected during 2007 and 2008, exept for the the FCI data from U1, which were accumulated from 2003 to 2007. At U1, the pre- and post-test data with physics concept tests have been collected for almost a decade. On the basis of the data, there are no significant variations among students from different years (*S11*). The BEMA and LCTSR data

were all taken in random sections within the same pool of population; therefore, these samples are representative of the students in the university.

Table S3. Summary of data from U.S. universities.

Samples	LCTSR			BEMA			FCI		
	Mean	*N*	SD	Mean	*N*	SD	Mean	*N*	SD
U1	76.5%	646	17.4%	31.0%	235	9.5%	49.4%	2592	19.3%
U2				24.1%	415	9.4%			
U3	75.8%	207	16.1%				44.6%	89	18.0%
U4	65.6%	207	18.9%						
Average of sample mean scores		72.6%			27.5%			47.0%	
Population mean*		74.2%			26.6%			49.3%	
Population SD		17.9%			10.0%			19.3%	

*In the paper, the population mean scores are used for comparison.

Table S4. Summary of data from Chinese universities.

Samples	LCTSR			BEMA			FCI		
	Mean	*N*	SD	Mean	*N*	SD	Mean	*N*	SD
C1				68.2%	120	11.9%	88.1%	182	11.1%
C2	74.0%	247	16.2%	64.1%	211	13.1%	85.0%	212	13.9%
C3	75.5%	85	14.7%				87.5%	122	10.1%
Average of sample mean scores		74.8%			66.1%			86.9%	
Population mean*		74.4%			65.6%			85.9%	
Population SD		15.9%			12.7%			12.1%	

Among the universities, U1 and C2 are the two in which all three tests have been given to the same population. Usually students in a single class took only one test and different classes were selected randomly to take different tests. The average class size is about 100.

In C2, we were able to collect data with all three tests in one class, which allows us to compute the correlations between students' scores on the different tests. We were also able to test one U.S. class in U1 with both FCI and LCTSR. The correlations are given in Table S5.

Table S5. Correlations between test results on LCTSR, FCI, and BEMA

Classes	LCTSR − FCI	LCTSR − BEMA	FCI − BEMA
C1 ($N = 80$)	0.12	0.17	0.70
U1 ($N = 102$)	0.23		

The results show a high correlation between FCI and BEMA and small correlations between LCTSR and FCI/BEMA. The interpretation of the correlation has to be carefully thought out. In this research, we study the possible causal interactions between science content learning and scientific reasoning. If a causal connection exists, a significant correlation between measures of content knowledge and reasoning is always expected. However, if no causal connection is evident, we may still observe correlations between measures of content knowledge and reasoning which may be the result of a wide variety of factors such as certain filtering and selection processes in the education system.

Comparisons with Results in the Literature and from Different Populations

To put our research results in context at the national level, we collected information from the literature and obtained additional data. The results are from freshmen college students in science and engineering majors enrolled in calculus-based introductory physics courses.

First, we discuss the U.S. population. FCI pretest scores are typically around 45% (*S12*). Additional data from another large state university similar to U1 give an average pretest score of 46% ($N = 355$, SD = 19%). BEMA pretest scores are typically around 26% (*S6*). We also had additional data from U2, which give an average pretest score of 25% ($N = 1631$, SD = 10%). We find our data (Table S3) to be comparable with results reported in the literature (*S12–S15*).

We have not identified much information in the literature about large scale LCTSR results for the studied population. From our data, we have observed significant differences between science/engineering majors enrolled in calculus-based intro-level physics courses and non-science/engineering majors in algebra based intro-level science courses (such as biology, chemistry, and physics). The results are summarized in Table S6.

In general, the test results of Chinese students from different universities are very similar. They have all received identical curricula in physics and must perform well on the same national college admission examination. Therefore, we consider the results in Table S4 to be representative of students in similar universities in China. There is also little published information on large scale assessment data using the three tests with Chinese college students. In addition, we found similar differences between the science/engineering majors and the non-science/engineering majors in China. From C2 and C3, we tested three classes of non-science/engineering majors with LCTSR. The average score was 58.1% ($N = 175$, SD = 20.1%), which is similar to that of the corresponding U.S. population (see Table S6).

Table S6. LCTSR test results of different U.S. populations. For non-science and non-engineering majors, there were 15 classes from U1 and U3 ($n = 1046$), and for science and engineering majors, there were 9 classes from U1, U3, and U4 ($n = 1061$). Averages or means ± SD.

Measure	LCTSR test results in the U.S.A. (%)	
	Non-science/engineering majors in algebra-based science courses	Science/engineering majors in calculus-based physics courses
Average of sample means	59.4 ± 3.8	72.0 ± 7.4
Population mean	59.0 ± 19.8	74.2 ± 16.2

Teaching Methods for Improving Scientific Reasoning

(*This part is incorporated from the earlier version of the supporting online materials*
http://www.sciencemag.org.proxy.lib.ohio-state.edu/cgi/data/323/5914/586/DC1/1.
The current version is at http://www.sciencemag.org.proxy.lib.ohio-state.edu/cgi/data/323/5914/586/DC1/2.)

Relations between intructional methods and the development of scientific reasoning has been widely studied (*S16-S20*). It is well documented that inquiry based science instruction promotes scientific reasoning abilities (*S17-S20*). Controlled studies have shown that students had higher reasoning abilities in inquiry classrooms versus non-inquiry classrooms (*S20*).

However, many of the existing studies are conducted with middle school and high school students. In order to see if the method also works in large scale college level scinece courses, we have conducted a plilot study at U4. The LCTSR was given as pre- and the post-test to students taking general science courses taught with different methods. We selected a student population that has a LCTSR pre-test score of 60.0% (N=263, SD=18.8%) in order to reduce the possible ceiling effect on the pre- and post-test changes. The population consists of students who have selected majors in a wide variety of fields but mechanical/electrical engineers and physical science majors are not included in this population.

Two one-semester courses are used in this study: Course A is an intro-level biology course taught in large lecture sessions with traditional lecture-laboratory method. Course B is a general science course taught in small classes (24 students max) with inquriy-based teaching method modeled after the Physics by Inquiry (*S21*). Course B curriculum contains diverse content topics including biology, chemistry, and earth science. All courses are algebra-based with very limited requirement on mathematics. The results are summarized in Table S7.

Table S7. LCTSR pre- and post-test results of different courses.

	Course A			Course B		
	N	Mean	SD	N	Mean	SD
Pre-Test	205	60.5%	19.3%	58	58.1%	17.3%
Post-Test	197	61.5%	19.9%	58	66.1%	16.0%
Pre-Post Difference		1.0%			8.0%	
Pre-Post Effect Size		0.05			0.47	

The data is consistant with the previous research, showing that inqury-based instruction can make a sizable change to student scientific reasoning ability in just one semster. This is an encouraging outcome; however, more research is needed in order to make it a solid finding and to understand the underlying mechanisms of the result.

References

S1. R. M., Hazen, J. Trefil. *Science Matters: Achieving Scientific Literacy*. New York: Anchor Books(1991).

S2. R. N. Giere, J. Bickle, R. F. Mauldin, *Understanding Scientific Reasoning*, 5th edition, Belmont, CA: Thomson/Wadsworth(2006).

S3. C. Zimmerman, *The Development of Scientific Reasoning: What psychologists contribute to an Understanding of Elementary Science Learning*. Paper commissioned by the National Academies of Science (National Research Council's Board of Science Education, Consensus Study on Learning Science, Kindergarten through Eighth Grade) (2005). http://www7.nationalacademies.org/bose/Corinne_Zimmerman_Final_Paper.pdf

S4. A. E. Lawson, "The development and validation of a classroom test of formal reasoning," *Journal of Research in Science Teaching* **15**(1), 11-24 (1978). Test used in study: Classroom Test of Scientific Reasoning, revised ed. (2000).

S5. A. Boudreaux, P.S. Shaffer, P.R.L. Heron, L.C. McDermott, "Student understanding of control of variables: Deciding whether or not a variable influences the behavior of a system," *American Journal of Physics* **76**(2), 163-170 (2008).

S6. A.E. Lawson, "The generality of hypothetico-deductive reasoning: Making scientific reasoning explicit," *The American Biology Teacher* **62**(7), 482-495 (2000).

S7. K. Cramer, T. Post, "Proportional reasoning," *Mathematics Teacher* **86**(5), 404-407 (1993).

S8. Z. Chen, D. Klahr, All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development* **70**, 1098–1120 (1999).

S9. D. Kuhn, D. Dean, Connecting scientific reasoning and causal inference. *Journal of Cognition and Development* **5**, 261–288 (2004).

S10. R. K. Hambleton, H. Swaminathan, *Item Response Theory: Principles and Applications*, Norwell, MA: Kluwer Academic Publishers (1985).

S11. L. Ding, N. W. Reay, A. Lee, L. Bao, "The effects of testing conditions on conceptual survey results," *Phys. Rev. ST Phys. Educ. Res.* **4**, 010112 (2008).

S12. Technical report at http://modeling.asu.edu/rup_workshop/resources/Project_Findings.pdf.

S13. S. J. Pollock and N. D. Finkelstein, "Sustaining educational reforms in introductory physics," *Phys. Rev. ST Phys. Educ. Res.* **4**, 010110 (2008).

S14. R. R. Hake, "Interactive-engagement vs traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," *American Journal of Physics* **66**, 64- 74 (1998).

S15 Also see Hake's website (http://www.physics.indiana.edu/~hake/) for more information on Hake's results of physics concept tests.

S16 Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, 27, 172-223.

S17 P. Adey, and M. Shayer, "Accelerating the development of formal thinking in middle and high school students," *Journal of Research in Science Teaching*, 27, 267-285, (1990).

S18 A. E. Lawson, *Science Teaching and the Development of Thinking*, Belmont, CA: Wadsworth Publishing Company, (1995).

S19 R. Benford, & A. E. Lawson, "Relationships between Effective Inquiry Use and the Development ofScientific Reasoning Skills in College Biology Labs," MS Thesis, Arizona State University. ERIC Accession Number: ED456157, (2001).

S19 E. A. Marek, & A. M. L. Cavallo, *The Learning Cycle and Elementary School Science*, Portsmouth, NH: Heinemann, (1997).

S20 B.L. Gerber, A.M. Cavallo, & E.A. Marek, "Relationships among informal learning environments, teaching procedures and scientific reasoning ability," *International Journal of Science Education*, 23(5):535-549, (2001).

S21 L. C. McDermott, *Physics by Inquiry*, Volumes I & II, John Wiley & Sons, Inc., New York, (1996).