

A Comparative Evaluation of STEM Education Indicators in the Era of Accountability

Jianjun Wang^{1,*}

¹Dept. of Advanced Educational Studies, California State University, Bakersfield, 9001 Stockdale Highway, Bakersfield, CA 93311, USA

*Correspondence: Tel: 1-661-654-3048 E-mail: jwang@csb.edu

Received: October 10, 2013 Accepted: November 25, 2013 Published: February 7, 2014

doi:10.5296/ije.v6i1.4406 URL: <http://dx.doi.org/10.5296/ije.v6i1.4406>

Abstract

As Common Core Standards gain momentum in the United States, it has been claimed to have curricular support from international studies. Accordingly, evidence of student learning is reviewed in this article to assess effectiveness of imported school curricula in the Washington DC area. The analysis is expanded to include confounding factors that impact school hours as a quantifiable variable. Meanwhile, additional indicators have been incorporated from higher education to facilitate the result triangulation across K-16 education. The research findings reveal importance of test fairness from comparative studies, i.e., when schools are under local control, no single international test can ensure a fair match to various curricula, nor does the result provide a valid measure of school accountability in STEM education.

Keywords: measurement issue; STEM education; assessment indicator

1. Introduction

Rapid development of information technology has transformed the landscape of education, and facilitated evidence gathering to support comparisons of student learning in a cross-national context. Accompanied with this change is a challenge of interpreting education indicators from different countries. Bonnet (2002) observed,

Policy makers show strong interest in international studies and comparative indicators because of the widespread belief that education is an investment necessary for the development of human capital and that there is a direct relationship between how good an education system is in terms of results and how successful the corresponding country is from an economic point of view. (p. 388)

As the global economy becomes more dependent on international competitions, quality of Science, Technology, Engineering, and Mathematics (STEM) education needs to be examined in comparative studies. A purpose of this research is to assess indicators of comparative education that impact curricular setting in local schools. As Arne Duncan, U.S. Secretary of Education, suggested that "To the extent that the U.S. can copy or adapt, and beg, borrow and steal successful practice from other nations, we should do so" (see White, 2012).

Built on the assumption that countries can learn from each other, research questions that guide this investigation are:

1. What are the evidences from the measurement of student learning to support implementation of new school curricula imported from other nations?
2. What are the confounding factors of school setting that impact the connection between school hours and education expectations?
3. What are the comparative indicators pertinent to assessing the leading position of U.S. in higher education?
4. What are the persistent issues of international testing directly hindering a fair comparison of student performance in a cross-national context?

In contrast to most action research within local schools, "International comparisons expand the range of comparison beyond the parochial limits of the U.S. national experience" (Commission on Behavioral and Social Sciences and Education, 1990, p. 2). Therefore, research in comparative education holds the promise of informing school accountability measures in the United States.

2. Literature Review

Educational accountability has gained attention of the general public for many years. Under George W. Bush's administration, No Child Left Behind Act of 2001 (NCLB) mandates annual testing in math and reading for all students in grades 3-8 and, at least once more, in grades 9-12 (Matthews, 2013). After President Obama entered the White House, more tests

have been added to the *Race to the Top* initiative (Acharya, 2013). The so-called “Accountability Movement” can be tracked back to the Cold War era when indicators of STEM education were linked to U.S. national security (Gibbs & Howley, 2000).

Despite U.S. economic recession since 2007, curriculum changes have been promoted by professional organizations to raise STEM education standards (White, 2012). In particular, the National Governors Association Center for Best Practices (NGA Center) and the Council of Chief State School Officers (CCSSO) released a report on June 1, 2009 to announce that 49 states and territories joined the Common Core State Standards Initiative (Tienken, 2010). Woolard (2013) observed, “Rigorous, common, and internationally competitive expectations are a key component of Race to the Top and accordingly, the vast majority of states signed on to implement a voluntary set of national standards—the Common Core” (p. 616).

According to Coleman, Pimentel, and Zimba (2012), “The Common Core State Standards (CCSS) were developed through an unprecedented, state-led initiative that drew on the expertise of teachers, parents, administrators, researchers and content experts from across the country” (p. 9). As an unprecedented event, Tienken (2010) suggested examination of evidences in student learning to support the curriculum change. He cautioned that results from international testing could be biased against U.S. students because “participating countries negotiate the test items” (p. 16). For example,

Representatives from more than 60 countries negotiated the development, wording, skills, and context of the items in the PISA 2003 assessment. The United States does not do well in these negotiations, as reflected by the fact that there are typically more test items covered by Asian curricula than by typical U.S. curricula. The Asian countries have reputations for scoring high on international tests, and there is intense national pressure to maintain that reputation. Their negotiating teams do their best to ensure that trend continues. (Tienken, 2010, p. 16)

In addition to the issue of curriculum coverage, confounding factors have been found to play profound roles in school settings (Osborn, 2004). More specifically, individual differences, as acknowledged by NCLB, may cause variations of content coverage in K-12 education. As Schmidt and Burroughs (2013) predicted, “If the ambitions of the Common Core initiative are realized, for the first time almost every public school student in the United States will be exposed to roughly the same content, especially in grades 1–8” (p. 54). Hence, contextual examination is needed to assess implication of the curriculum change on school hours and education expectations.

Furthermore, the impact is not confined within K-12 levels. King and Jones (2012) noted, “A national movement is bringing educators together from across the Pre-K to 16 continuum with a common goal: re-establishing the United States as an international leader in educational attainment” (p. 16). Since no international testing was conducted beyond high school, alternative indicators should be explored to assess the leading position of U.S. in higher education. Because the media exposure of comparative findings has often been oversimplified, the K-16 approach can help enrich understanding of comparative findings. As Loveless (2011) pointed out, “Whether commentators perpetrating myths of international

testing, states winning races while evidencing only mediocre progress, or an eighth-grade test dominated by content below the eighth grade, the story is rarely as simple as it appears on first blush” (p. 1).

Digesting results from comparative studies has become an inescapable task for U.S. educators because “The [Common Core] Standards also draw on the most important international models” (Common Core State Standards Initiative (2010, p. 3). According to Schmidt and Burroughs (2013), “The adoption of the Common Core State Standards by 46 states and the District of Columbia represents a dramatic departure in U.S. education” (p. 54). In the past, McPartland and Schneider (1996) asserted that “the change in public policy that is aimed at giving all students access to a common core curriculum of high standards will fail to reach its full potential in its current form, which largely ignores the issues of students' diversity” (p. 78). Hence, technical issues from international testing need to be reviewed to evaluate the impact of curricular changes for U.S. students from diverse background.

In summary, research literature on Common Core Standards (CCS) has informed this study on multiple fronts. Because of its consideration of international models, investigations are needed to review past evidences of student learning in local settings (Question 1). In addition, the new curriculum design is inseparable from consideration of school hours and education expectations (Question 2). “Despite the critical need for collaboration between K-12 and postsecondary education to ensure that students graduate from high school ready to succeed in college, the two sectors have traditionally maintained an arm's length relationship” (King & Jones, 2012, p. 16). Hence, Question 3 has been adduced for this investigation to articulate indicators of STEM education beyond high school. Question 4 needs to be examined because CCS demands that “all states voluntarily adopt the same set of curriculum standards and, eventually, submit to a national test” (Tienken, 2010, p. 14). In that context, persistent issues of international testing should be reviewed to overcome similar issues in the national testing. Altogether, the literature review has summed up past studies to position the four research questions for this study in broad-based inquiries of comparative education.

3. Methods

In searching for quick fixes, politicians tended to rely on simplistic solutions that were quantifiable in nature. Consequently, Osborn (2004) observed,

In recent years there has been a growing tendency to 'borrow' educational policies and practices from one national setting where they appear to be effective and to attempt to transplant these into another, with little regard for the potential significance of the cultural context into which they will be imported. (p. 265)

To address Question 1, a cultural comparative approach has been taken to examine evidences of student learning from Singapore curricula imported to Washington, DC, and the analyses are expanded to include perspectives of different countries in a cross-cultural context.

School hour was a quantifiable measure subjected to international comparisons in the research literature, but the time configuration was often based on class schedules (Wray, 1999). Michael Sadler's (1900) classic quote already indicated that “in studying foreign systems of education we should not forget that the things outside the schools matter even more than the things inside the schools, and govern and interpret the things inside” (p. 310). Since extracurricular activities play an important role in college admission, Question 2 is examined through a comparison of external factors that govern student learning inside school.

King and Jones (2012) further stressed that “Higher education institutions and schools must collaborate to define the knowledge and skills students need to be ready for college and to determine the most effective means of helping students meet those expectations” (p. 16). As K-16 enrollments increased over past three decades, a systematic approach has been taken to address Question 3, and additional indicators are derived from postsecondary education.

It should be noted that “When it comes to method, the majority agreement is that there is no methodology peculiar to comparative research” (Heidenheimer, Heclo, & Adams, 1983, p. 505). Given the method variation, result triangulations have been made to scrutinize persistent issues of international testing that undermine fairness of the measurement outcomes in STEM education (Question 4). Best and Khan (2005) maintained that triangulation was an effective method to verify and validate data for educational comparisons.

In summary, an analytical framework has been identified in this study to support *fact findings* (Questions 1 & 2), *quality assessment* (Question 3), and *result triangulation* (Question 4). As Jowell (1998) pointed out, “The strict standards we apply to the evaluation of national surveys are too often suspended when it comes to cross-national studies” (p. 168). This study has been designed to tackle the weakness of research accountability in comparative education.

4. Evidences from Implementing A Foreign Curriculum in the U.S.

In reporting the 2007 results from Trend in Mathematics and Science Study (TIMSS), Hanushek and Peterson (2011) noted that “Twice as many students in Singapore are proficient in math as in the United States” (p. 1). Thus, Singapore curricula have been adopted in approximate 2,000 U.S. schools (Turque, 2011). Bruce-Monroe School participated in that trial at Washington, D.C. To evaluate the impact on student learning, the school has gone through repeated measurements with striking results highlighted in Washington Post:

On the 2009 D.C. Comprehensive Assessment System, the first citywide test after the changeover, pass rates remained virtually unchanged, with 49 percent of students achieving proficiency. Last year, the pass rate at the school plunged to 23 percent. That decline was steeper than a citywide drop. (Turque, 2011, p. 10)

Apparently, the Singapore approach did not demonstrate a better result from the U.S. experiment, nor did Singapore sustain its leading position in student performance. In 2009,

the Organisation for Economic Co-operation and Development (OECD) administered a worldwide Programme for International Student Assessment (PISA) to measure how well a nation's education system prepared its students for the global economy. It was reported that

Nations such as South Korea, Finland, and Singapore have traditionally topped the rankings, but, apparently, even they are no match for Shanghai, which shoved the others into lower positions in its very first year of participation in the programme, in 2009. (Jiang, 2011)

In a quick reaction, the New York Times acknowledged that “the stellar academic performance of students in Shanghai was noteworthy, and another sign of China’s rapid modernization” (Dillon, 2010, p. 4). On basis of the new evidence, one might wonder whether politicians have contemplated a new policy to replace the Singapore curricula with the national curricula of China.

In reviewing the results from comparative studies, however, Baker (2007) argued that “for the U.S. and for the top dozen or so most-advanced nations in the world, standings in the league tables of international tests are worthless” (p. 101). More specifically, schools in the U.S. are under local control. No single test from any international measurements can provide a good match to various curricula across the nation (Keitel & Kilpatrick, 1999). If politicians really wished to have a quick fix, legislative actions could have been taken to adopt a national curriculum like the ones from China and Singapore, and use it to regulate teaching practices across all schools in the U.S. But this “one size fit all” approach will inevitably ignore variation in local needs and individual developments.

Instead of assessing what students learned at a particular school in the past, Scholastic Aptitude Test (SAT) scores have been considered in U.S. college admission to predict individual learning ability from the future perspective. Based on the SAT evidence, American students clearly outperformed their peers from China. As Yang (2011) reported,

Chinese students are surpassed by their US counterparts at the Scholastic Assessment Test (SAT) needed to enter a US university, ... The Chinese test emphasizes memorization while the SAT consists of three 800-point sections - mathematics, critical reading and writing – focusing on independent thinking and creativity, which many US colleges deem as indispensable for academic studies. (p. 1)

Preparation of independent thinking and creativity often involves open-ended explorations (Cooper, 2006). In general, “Chinese students spend less time than American students on athletics, music and other activities not geared toward success on exams in core subjects” (Dillon, 2010, p. 26). The time commitment has skewed Chinese student learning toward preparation for one critical exam, i.e., College Entrance Examination (or “gaokao”). Under the test-driven culture, Chinese parents often pay for supplemental classes on Saturdays or during summer breaks. “When students reach pre-college classes, preparation for the gaokao becomes all consuming” (Hopper, 2010, p. 27). If so many extra hours have been devoted in China to cover the narrowly-delimited content of College Entrance Exam, schools in the U.S. would need much more time to cover its broad curricula that are “a mile wide”.

To pursue the depth of learning, Vygotsky (1997) suggested that educators should design teaching activities within “the zone of proximal development” (ZPD). Accordingly, instructional content should be positioned slightly ahead of student progress. In China, students who made a rapid progress will become the first ones to reach the strict curriculum boundary. For instance, calculus courses are offered at many high schools in the U.S. for advanced students. However, few Chinese students have this learning opportunity because of its irrelevancy to College Entrance Examination. If the PISA test included questions in calculus, students from Shanghai might not do so well in the international comparison.

Although education is desired to support transferring and expanding knowledge, skills, and values from one generation to another (Dewey, 1944), the dull drills in paper-and-pencil exercises have formed a passive learning habit to hinder that process. When taking classes in the U.S., Chinese students are often reluctant to contribute ideas in discussions (Harris, 2012). Similarly, most national examinations in developing countries require a high level of factual recall, but little in the way of important initial competencies or critical thinking skills (Somerset, 2011).

Despite the mediocre rankings of U.S. education in international measurement, Baker (2007) argued that “There is no association between test scores and national success” (p. 101). When students in east Asia outperformed their peers from the U.S. in the late 1990s, the United States had the world’s largest economy, but “the economy of Asia [was] in its worst state since the end of this Second World War” (Thorsten, 2000, p. 50). Meanwhile, grass looks greener on the other side of the fence. Chinese parents seemed to have puzzled out the value judgment between the two school systems, and preferred to send their single child to the U.S. for further education (Orson, 2012).

5. Confounding Factors in School Settings

As an education institution, school supports curriculum-based learning. Schmidt, McKnight, Cogan, Jakwerth, and Houang (1999) examined U.S. curricula in mathematics and science education, and found them to be “a mile wide and an inch deep”. Since it takes time to deepen the knowledge inquiry, U.S. leaders attempted to justify the need of more school hours through a comparison of instructional time across different education systems (see Wray, 1999).

In comparison to Asian education systems, such as those in China and Japan, American school years are at least one month shorter (Stewart, 2009; Wray, 1999). Accumulating across K-12 grades, American students seem to have lost more than one year of schooling before college admission. It was further claimed that an extra year in education has a 4-7% impact on gross domestic product (Bonnet, 2002). For this reason, President Obama indicated that “I know longer school days and school years are not wildly popular ideas, ... But the challenges of a new century demand more time in the classroom” (see Quaid, 2011, p. 25).

As a quantifiable indicator, the time configuration can be compared across nations. In reality, however, school hours only count for a small portion of student daily life. Walberg (1984) reported that children from birth to age 18 spent 90% of their time outside of school. Therefore, it might be farfetched to link the school hours with the entire student learning experience. Additional time on extracurricular activities has been found important for students to explore creative projects (Emmerson, 2012). In China, the precious learning time has been absorbed during the poor mining of limited knowledge for College Entrance Exam (Flores, 2011), and thus, “most of those high scorers rarely end up making history by becoming the next Bill Gates or Mark Zuckerberg” (Hopper, 2010, p. 2)

To facilitate well-rounded student development, high schools in the U.S. often support various student clubs during the extracurricular hours. President Obama announced an agenda to expand those high-quality afterschool programs (see Talbott, 2008). In addition, the U.S. education system allows concurrent enrollment of high school students at the college level. This policy supports a good use of extracurricular hour to dig into the rich mine of knowledge beyond the level of secondary education (Robertson, Chapman, & Gaskin, 2001). High schools also allow students to take advanced placement (AP) courses or enroll in an international baccalaureate program to earn college credits. Researchers found beneficial effects from these early start programs, regardless of student ethnic backgrounds (Karp, Calcagno, Hughes, Jeong, & Bailey, 2007).

The value of extracurricular activities could have been diminished if school periods were blindly prolonged in the U.S. according to the Chinese model. School creativity could have been suppressed by reckless government regulations. Under the banner of increasing educational accountability, the George W. Bush administration attempted to prescript instruction using No Child Left Behind (NCLB). However, as White (2012) reported, “Since NCLB was implemented in 2002, American performance has actually decreased. American students ranked 18th in mathematics in 2000 but fell to 31st place in 2009, according to OECD data” (p. 4).

In retrospect, the school periods in the U.S. are almost identical between secondary and tertiary education, but colleges have been rarely criticized for having an insufficient number of school days. Interpretation of these seeming quantifiable indicators demands an in-depth examination of qualitative factors, including accommodation of curricular differences, extracurricular activities, and pressure for college admission.

President Obama announced that “Jobs today often require at least a bachelor's degree”, and “the unemployment rate for folks who've never gone to college is over twice as high as for folks with a college degree or more” (see Huffington, 2010, p. 3). Since the human resource development is not confined within compulsory education, it seems appropriate to divert some attention to quality indicators of higher education that have more profound impact on the future economy.

6. Quality Indicators in Higher Education

Unlike test score comparisons at the secondary level, no international studies have been conducted to test student performance in higher education. When reporting education quality, local media often neglected college students. Whereas most U.S. citizens have gone through compulsory education, higher education is not free for everyone, nor does the college campus reside in every community.

To broaden the horizon on postsecondary education, Rotberg (1991) suggested five dimensions for assessing quality of U.S. higher education:

1. How productive is the U.S. in basic and applied research fields? What does the marketplace say about the research opportunities in our institutions of higher learning?
2. What are our accomplishments in making major technological advances, as measured by patents and their application in products, in areas such as semiconductors, biotechnology, materials development, radiation imagery chemistry, information storage and retrieval, medical research, and pharmaceuticals?
3. Are the fields of science and engineering attracting high-achieving students? Is there a shortage of students or faculty members in these fields?
4. Does the teaching give students who do not major in those fields some understanding of key scientific issues and methods?
5. Are we maintaining the technical expertise of the workforce? (p. 300)

Tienken (2010) reported that “the United States has ranked either 1st or 2nd consistently in economic competitiveness since 1998. The only year the United States fell out of the top two spots was in 2006, following Hurricane Katrina” (p. 17). As the country maintains its leading position in basic and applied research fields, it also provides great learning opportunities to attract high-achieving students and scholars around the world. As William Wulf (2005), President of U.S. National Academy of Engineering, pointed out, “Between 1990 and 2004, over one third of Nobel Prizes in the United States were awarded to foreign-born scientists. ... Top-notch students and teachers from abroad help make U.S. colleges and universities global centers of excellence and diversity” (p. 2).

In comparison, China has the second largest economy in the world. However, higher education in the U.S. is far more attractive for foreign students. After entering the 21st century, 37% of the terminal degrees from U.S. universities were awarded to foreign students (Cao, 2008). More recently, “Not only is China the largest country of origin for international graduate students in the U.S., but its rate of growth is far outpacing all other countries and regions in the survey, including South Korea and India” (Korn, 2012, p. 2). Thus, U.S. higher education is doing well on the first three dimensions.

Regarding the fourth dimension, Redish and Steinberg (1999) observed that “Many physics faculty come away from teaching introductory physics deeply dismayed with how little the

majority of their students have learned” (p. 24). This is because the priority of mathematics or science department typically places scientific research first, followed by the education of Ph.D. students, then Masters students, upper-level undergraduate majors and courses, introductory courses for majors, introductory courses for subject-related professionals, and finally the lowest of the low and often entirely absent, science/mathematics for non-specialists (Burnside, 2002). As a result, “Pedagogical theory is generally held in low esteem by university scientists” (Hestenes, 1987, p. 440).

With insufficient emphasis on educational research to streamline the knowledge accumulation in STEM teaching, widespread student learning issues often need repeated discoveries of similar solutions through trial-and-error methods in a disorganized manner. For instance, Su, Su, and Goldstein (1994) reported a study of Chinese teachers who observed performance of U.S. teachers in a K-12 setting. The result was astonishing – “Several Chinese scholars witnessed American teachers and students carrying out the wrong scientific experiments and calculations with great enthusiasm” (Su, Su, & Goldstein, 1994, p. 260).

The United States did not seem so competitive on the fourth dimension because “Chinese math and science teachers, at least in urban areas, receive more-rigorous training than their U.S. counterparts” (Cavanagh, 2006, p. 7). President Obama attempted to improve teacher quality by linking teachers' pay to student performance (Klein, 2009), but the issue could have been better solved at a pre-service stage when teacher candidates are taking core courses in mathematics and science departments.

Downplaying the role of teaching has led more Americans to avoid the fields of STEM education (Joyner, 2011), which will reciprocally hurt student pipeline preparation in the United States (i.e., the fifth dimension). But the concern is not so severe from the standpoint of higher education. The globalization has allowed the U.S. to attract high-achieving students from foreign countries. For instance, Mong (2012) observed that “more [Chinese] students opt out of the gaokao and sign up for exams like the TOEFL (Test of English as a Foreign Language) and the SAT (Scholastic Aptitude Test), both of which are generally prerequisites for applying to any U.S. college and university” (p. 25). Because the majority of Americans are descended from someone who arrived from another country, technical issues of international assessments should be further examined to improve comparative evaluation of education quality between U.S. and other countries.

7. Practical Issues in Comparative Measurement of Education Quality

U.S. News & World Report teamed up with American Institutes for Research (AIR) to evaluate quality of 21,776 public schools across the United States (Morse, 2012). School quality has been identified through a three-step approach: (1) Determine whether student performance at a particular school is above the average student performance in the state; (2) “For those schools that made it past this first step, the second step determined whether the school's least-advantaged students (black, Hispanic, and low-income) were performing better than average for similar students in the state” (Morse, 2012, p. 6); (3) College readiness is

examined according to student test performance from Advanced Placement or International Baccalaureate tests, both having the content-based rigor to earn college credits for high school students.

These practical steps have generated more credible results for education stakeholders, such as parents and teachers, to make a school choice. (Knorr, 2010). Nonetheless, the useful guidance in the U.S. has been contradicted by comparative results from international measurement. More specifically, Beverly Hills Unified School District (BHUSD) is one of the top performing districts in the United States. While the district's students scored in the 80th percentile in recent state and national reading and math assessments, the position dropped to the 53 percentile in both subjects when comparing to countries such as Finland, China, Korea and Canada (White, 2012).

While results from U.S. News & World Report have been found useful by local educators in the U.S. (see Sheehy, 2013), credibility of international testing has been constantly challenged by researchers (Jowell, 1998; Tienken, 2010; Wang, 2001). For illustration, measurement features of international testing are examined below to discuss three technical issues impacting a fair comparison of student performance across countries:

7.1 Twist of Item Difficulty

In the domestic project on school quality measurement, U.S. News and World Report incorporated “reading and math results for all students on each state's high school proficiency tests” (Morse, 2012, p. 5). While mathematics is considered as an international language (Jacobs, 2010), reading is a basic skill embedded in the testing process. Despite the assiduous effort on instrument translation to support comparative measurements (e.g., O'Connor & Malak, 2000), “incorrect responses were not necessarily an indication of a lack of knowledge of a concept; sometimes they were due to misinterpretation of a question, a word, a phrase or a diagram” (Harlow & Jones, 2004, p. 234). This is because reading comprehension often depends on local contexts, and the same English phrase could be interpreted differently in foreign countries (Dowd, 2012). For instance, the following item was included in the TIMSS 2003 instrument.

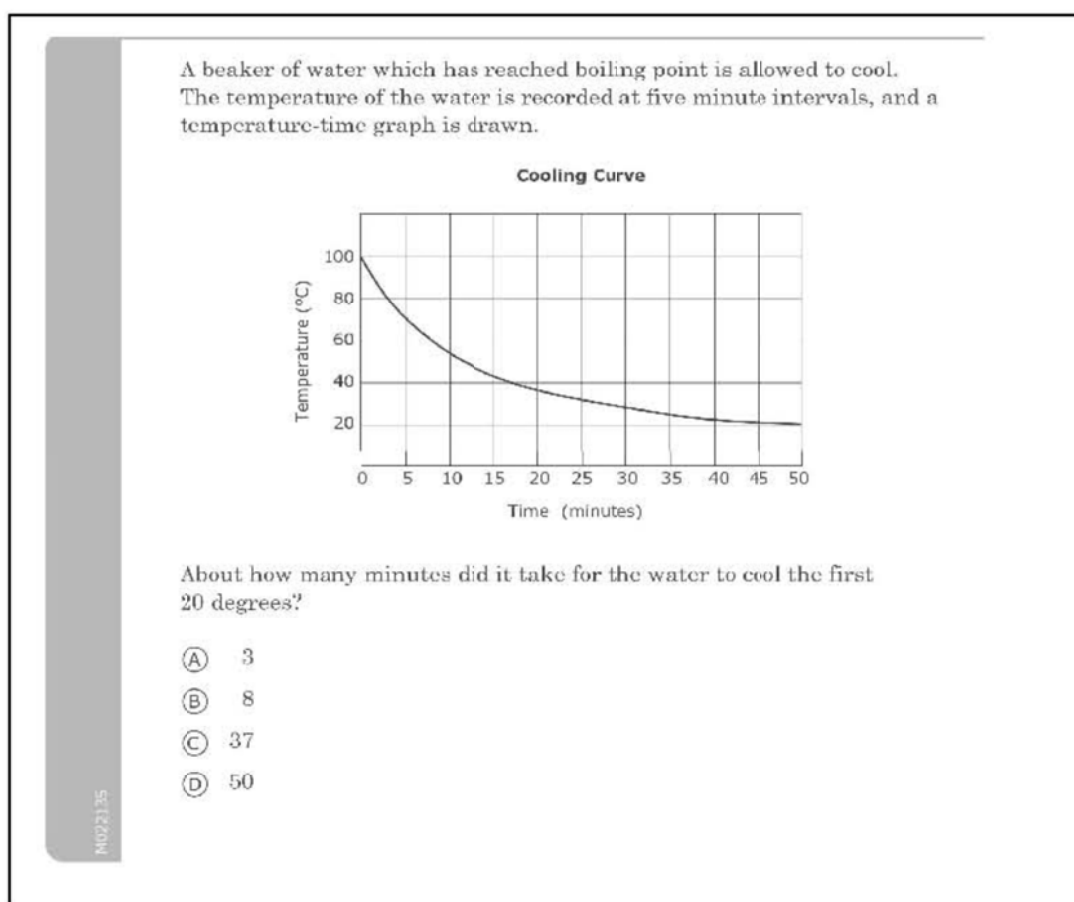


Figure 1: M022135 of the TIMSS 2003 Study

The phrase “the first 20 degrees” was not commonly used in Hong Kong. Omitting the explicit words “the first”, it would typically be said as “to cool 20 degrees”. As Harlow and Jones (2004) pointed out, many students in Hong Kong misread the question as “to cool to 20 degrees” instead of “to cool the first 20 degrees”. Such a language barrier has twisted the item difficulty, causing an unfair score comparison for Hong Kong students.

7.2 Neglect of Heterogeneity in Student Performance

Diversity of student population is interrelated with school quality (U.S. Department of Education, 2004). Thus, the U.S. News and World Report incorporated indicators of school effort in helping least-advantaged students (Black, Hispanic, and low-income). Sloane (2008) suggested that “We change the basic research question from what works to what works for whom and in what contexts” (p. 43). Because the United States is the only developed nation that has a growing population and is becoming increasingly diverse (Crouch, 2012), not all countries share the same interest in examining the heterogeneity of student performance on the ethnicity dimension, nor has the poverty level been mindfully assessed in comparative studies (Clarke et al., 2008).

Nonetheless, heterogeneity of the student population is undeniably reflected in test scores of international measurement. In examining the results from the first round of TIMSS in 1995

and a repeat of the TIMSS (TIMSS-R) in 1999, Berliner (2001) pointed out,

Average scores mislead completely in a country as heterogeneous as ours . . . In science, for the items common to both the TIMSS and the TIMSS-R, the scores of white students in the United States were exceeded by only three other nations. But black American school children were beaten by every single nation, and Hispanic kids were beaten by all but two nations. A similar pattern was true of mathematics scores. (p. B3).

Without including poverty and ethnicity factors in the sample stratification, international measurements cannot differentiate comparative findings for the least-advantaged students (Black, Hispanic, and low-income).

7.3 Oversight of Guessing Effect in Instrument Designing

The U.S. News and World Report assessed college readiness according to student test performance from Advanced Placement (AP) or International Baccalaureate (IB) tests. Those tests are solely discipline-based with strict rigor to justify college credits for high school students. In contrast, “Difficulty in international comparison hinges on considerable uncontrolled variation in variables other than those of policy interest” (Commission on Behavioral and Social Sciences and Education, 1990, p. 3).

Although TIMSS items were endorsed by a Subject Matter Advisory Committee that included “distinguished scholars from 10 countries” (Beaton et al., 1996, p. A-9), the item design did not ensure accurate results. For instance, one item labelled S022281 in the TIMSS 2007 project read,

A tray containing 300 grams of water is placed in the freezer to make ice. What is the mass of the ice after the water freezes?

(Check one box.)

More than 300 grams

300 grams

Less than 300 grams

Explain your answer.

The second box could be chosen as a correct answer when water evaporation was negligible during the freezing process. If the evaporation effect was included, the ice mass could be “Less than 300 grams”.

Since no chemical reaction occurred, it would be impossible to obtain more water from nowhere, and thus, “More than 300 grams” of ice is a wrong choice. Nonetheless, the TIMSS 2007 grading rubrics treated the first box as an acceptable choice as well. The explanation was based on a pretext of trapped air within the ice!

The wrong rubric could have caused misconceptions in science teaching. If the ice mass were to increase without adding more water, that mechanism would be used to irrigate desert and

improve the global environment! Because TIMSS researchers have made all three check boxes admissible, this item not only imbedded a scientific error, but also skewed the results of international comparison by promoting blind guessing.

In summary, comparative assessments of student learning are much more complex than simple score reporting. Examples in this section illustrated needs for improving international STEM testing on at least fronts: (1) Avoid inadvertent twist of item difficulty in the cross-cultural context, (2) Establish discriminant validity to disentangle heterogeneity of student performance, and (3) Correct oversights of scientific errors and minimize the impact of random guessing.

8. Conclusion

When the Common Core Standards were advocated, one important rationale was to support U.S. economic competitiveness through improvement of American student performance in international testing (Woolard, 2013). Before accepting the premise of having valid results from international testing, educators should be reminded that no single international test can provide adequate coverage for various curricula in the U.S. In generating quality indicators for STEM education, not all the test developers have demonstrated subject competency at the level of secondary education, and wrong items might have been incorporated to skew the results of international comparison. A hasty change of local curriculum may lead to a repeat of the lessons from importing Singapore curricula at Bruce-Monroe School in Washington, DC. Instead, more considerations should be given to domestic indicators that have demonstrated their usefulness in the past, including enhancement of teacher preparation and student inquiries in STEM education.

References

- Acharya, E. (2013). I'm proof that teaching to the test doesn't work. Retrieved from <http://www.policymic.com/articles/70517/i-m-proof-that-teaching-to-the-test-doesn-t-work>
- Baker, K. (2007). Are international test scores worth anything? *Phi Delta Kappan*, 89(2), 101-104.
- Beaton, A., Martin, M., Mullis, I., Gonzalez, E., Smith, T., & Kelly, D. (1996). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: Boston College.
- Berliner, D. (2001). *Averages that hide the true extremes*. Washington Post, Outlook Section, January 28.
- Best, J., & Kahn, J. (2006). *Research in education*. New York: Pearson.
- Bonnet, G. (2002). Reflections in a critical eye: On the pitfalls of international assessment.

Assessment in Education, 9(3), 387-399.
<http://dx.doi.org/10.1080/0969594022000027690a>

- Burnside, M. E. (2002). *Physics education research: Summation and application* [MS Thesis]. Santa Cruz, CA: University of California.
- Cao, C. (2008). Tsinghua and Peking university graduates “dominate” U.S. doctoral degrees. Retrieved from <http://hk.chem8.org/bbs/thread-17442-1-1.html>
- Cavanagh, S. (2006). China takes different tack from U.S. in teaching mathematics and science. *Education Week*, 25(41), 7.
- Clarke, D. J., Mesiti, C., O’Keefe, C., Xu, L.H., Jablonka, E., Mok, I. A. C., & Shimizu, Y. (2008). Addressing the challenge of legitimate international comparisons of classroom practice. *International Journal of Educational Research*, 46(5), 280-293. <http://dx.doi.org/10.1016/j.ijer.2007.10.009>
- Coleman, D., Pimentel, S., & Zimba, J. (2012). Three core shifts to deliver on the promise of the Common Core Standards. *State Education Standard*, 12, 9-12.
- Commission on Behavioral and Social Sciences and Education (1990). *A framework and principles for international comparative studies in education*. Washington, DC: National Academy Press.
- Common Core State Standards Initiative (2010). *Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects*. Washington, DC: Author (ERIC Reproduction Service No. ED522008).
- Cooper, J. (2006). *Classroom teaching skills*. New York: Houghton-Mifflin.
- Crouch, R. (2012). *The United States of education: The changing demographics of the United States and their schools*. Alexandria, VA: Center for Public Education.
- Dewey, J. (1944). *Democracy and Education*. New York: The Free Press.
- Dillon, S. (2010, December 7, U.S. section). Top test scores from Shanghai stun educators. Retrieved from <http://www.nytimes.com/2010/12/07/education/07education.html?pagewanted=all>
- Ditkoff, D. (2008). Not everything that counts can be counted; and not everything that can be counted counts. Retrieved from http://www.ideachampions.com/heart/archives/2008/05/now_everything.shtml
- Dowd, A. (2012). An NGO perspective on assessment choice: From practice to research to practice. *Compare*, 42, 541-545.
- Emmerson, J. (2012). Finding yourself in school: A literature review through the thematic lenses of identity and music. *Canadian Journal for New Scholars in Education*, 4, 1-9.
- Flores, C. (2011). Which superpower has the better education? (U.S. or China?) Retrieved from

<http://youthvoices.net/discussion/which-superpower-has-better-education-us-or-china>

- Gibbs, T., & Howley, A. (2000). "World-Class Standards" and local pedagogies: Can we do both? Charleston, WV: ERIC Clearinghouse on Rural Education and Small Schools. (ERIC Document Reproduction Service No. EDO-RC-008)
- Gonzalez, E.J., Galia, J., & Li, I. (2004). Scaling methods and procedures for the TIMSS 2003 mathematics and science scales. In M.O. Martin, I.V.S. Mullis, & S.J. Chrostowski (Eds.), *TIMSS 2003 Technical Report* (pp. 252-273). Chestnut Hill, MA: International Study Center, Boston College.
- Hanushek, E., & Peterson, P. (2011, August 28). Why can't American students compete? Retrieved from <http://www.thedailybeast.com/newsweek/2011/08/28/why-can-t-u-s-students-compete-with-the-rest-of-the-world.html>
- Harlow, A., & Jones, A. (2004). Why students answer TIMSS science test items the way they do. *Research in Science Education*, 34, 221–238. <http://dx.doi.org/10.1023/B:RISE.0000033761.79449.56>
- Harris, D. (2012). *Chinese student in America: It's bad out there*. Seattle, WA: Harris & Moure.
- Heidenheimer, A., Hecl, H., & Adams, C. (1983). *Comparative Public Policy*. New York: St. Martin's Press.
- Hestenes, D. (1987). Toward a modeling theory of physics instruction. *American Journal of Physics*, 55, 440-454. <http://dx.doi.org/10.1119/1.15129>
- Hopper, J. (2010). Is China's education system keeping up with growing superpower? Retrieved from <http://abcnews.go.com/WN/China/chinas-education-system-helping-hurting-superpower-s-growing-economy/story?id=12152255#.T8mYsFIvAk4>
- Huffington, A. (2010). *Third world America*. Retrieved from http://www.ontheissues.org/Archive/Third_World_America_Education.htm
- Jacobs, I. (2010). W3C integrates math on the web with MathML 3. Retrieved from <http://www.w3.org/2010/09/mathml-pr>
- Jiang, X. (2011). How Shanghai schools beat them all. Retrieved from <http://the-diplomat.com/2011/08/01/how-shanghai-schools-beat-them-all/>
- Jones, R. (2010). President's Council of S&T Advisors issues strategy to transform K-12 STEM education. Retrieved from <http://www.aip.org/fyi/2010/099.html>.
- Jowell, R. (1998). How comparative is comparative research? *American Behavioral Scientist*, 42, 168-177. <http://dx.doi.org/10.1177/0002764298042002004>
- Joyner, J. (2011). Why more Americans don't major in the math and science? Retrieved from

<http://www.outsidethebeltway.com/why-more-americans-dont-major-in-the-math-and-science/>

- Karp, M., Calcagno, J., Hughes, K., Jeong, D. W., & Bailey, T. (2007). The postsecondary achievement of participants in dual enrollment: An analysis of student outcomes in two states. Community College Research Center. Retrieved from <http://www.ecs.org/html/IssueSection.asp?issueid=214&s=Selected+Research+%26+Readings>.
- Keitel, C., & Kilpatrick, J. (1999). Rationality and irrationality of international comparative studies. In G. Kaiser, I. Huntley, E. Luna (Eds.) *International comparative studies in mathematics education* (pp. 241-257). London: Falmer.
- King, J. E., & Jones, A. (2012). The Common Core State Standards: Closing the school-college gap. *Trusteeship*, 20, 16-21.
- Klein, K. (2009, March 10). Obama introduces first part of US education reform plan. Retrieved from <http://www.voanews.com/english/2009-03-10-voa38.cfm>
- Knorr, R. (2010). High achievers: Sarasota, Manatee and Charlotte schools continue to excel. Retrieved from <http://www.florida-homebuyer.com/archives/sarasota/63/Annual-2008/724/High-Achievers.php>
- Korn, M. (2012). Chinese applicants flood U.S. graduate schools. Retrieved from <http://online.wsj.com/article/SB10001424052702304750404577319922446665462.html>
- Li, Y. (2009, April). *Understanding the scaling methodologies in large-scale assessments: A TIMSS case study*. Paper presented at 2009 AERA Annual Meeting, San Diego, CA.
- Loveless, T. (2011). *The 2010 Brown Center Report on American Education: How well are American students learning?* Washington, DC: Brookings Institution.
- Matthews, M. (2013). *Benefits and drawbacks of state-level assessments for gifted students: NCLB and standardized testing*. Retrieved from <http://tip.duke.edu/node/827>
- McPartland, J., & Schneider, B. (1996). Opportunities to learn and student diversity: Prospects and pitfalls of a Common Core Curriculum. *Sociology of Education*, 69, 66-81. <http://dx.doi.org/10.2307/3108456>
- Mong, A. (2012). *Chinese applications to U.S. schools skyrocket*. Retrieved from http://behindthewall.msnbc.msn.com/_news/2012/01/11/9679479-chinese-applications-to-us-schools-skyrocket?lite
- Morse, R. (2012). Best high school methodology: U.S. News looked at thousands of public schools to identify the most outstanding. Retrieved from <http://www.usnews.com/education/high-schools/articles/2012/05/07/best-high-schools-methodology>
- O'Connor, K., & Malak, B. (2000). Translation and cultural adaptation of the TIMSS

- instruments. In M.O. Martin, K.D. Gregory, K.M. O'Connor, and S.E. Stemler (eds.), *TIMSS 1999 Benchmarking Technical Report*. Chestnut Hill, MA: Boston College.
- Orson, D. (2012). International students enroll in US high schools. Retrieved from <http://www.yourpublicmedia.org/content/wnpr/international-students-enroll-us-high-schools>
- Osborn, M. (2004). New methodologies for comparative research? Establishing 'constants' and 'contexts' in educational experience. *Oxford Review of Education*, 30, 265-285. <http://dx.doi.org/10.1080/0305498042000215566>
- Quaid, L. (2011). President Obama is bad for public education. Retrieved from <http://jaxkidsmatter.blogspot.com/2011/03/president-obama-is-bad-for-public.html>
- Redish, E. F., & Steinberg, R. N. (1999). Teaching physics: Figuring out what works. *Physics Today*, 52, 24-30. <http://dx.doi.org/10.1063/1.882568>
- Rotberg, I. C. (1991). How did all those dumb kids make all those smart bomb? *Phi Delta Kappan*, 72, 788-781.
- Sadler, M. (1900). How can we learn anything of practical value from the study of foreign systems of education? In J. H. Higginson (Ed.), *Selections from Michael Sadler: Studies in world citizenship*. Liverpool: DeJall & Meyorre.
- Schmidt, W., & Burroughs, N. (2013). How the Common Core boosts quality & equality. *Educational Leadership*, 70(4), 54-58.
- Schmidt, W. C., McKnight, C., Jakwerth, P. R., & Houang, R. (1999). *Facing the consequences: Using TIMSS for a closer look at U.S. mathematics and science education*. Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Sheehy, K. (2013). U.S. News releases 2013 best high schools rankings. Retrieved from <http://www.usnews.com/education/high-schools/articles/2013/04/23/us-news-releases-2013-best-high-schools-rankings>
- Sloane, F. (2008). Through the looking glass: Experiments, quasi-experiments, and the medical model. *Educational Researcher*, 37(1), 41-46. <http://dx.doi.org/10.3102/0013189X08314835>
- Somerset, A. (2011). Strengthening educational quality in developing countries: The role of national examinations and international assessment systems. *Compare: A Journal of Comparative and International Education*, 41, 141-144.
- Stewart, V. (2009). China and U.S. can swap ideas about math and science. *Phi Delta Kappan*, 91(3), 94-95.
- Su, Z., Su, J., & Goldstein, S. (1994). Teaching and learning science in American and Chinese high schools: A comparative study. *Comparative Education*, 30, 255-270. <http://dx.doi.org/10.1080/0305006940300307>

- Talbott, J. (2008). Obamanomics. Retrieved from http://www.ontheissues.org/Archive/Obamanomics_Barack_Obama.htm
- Thorsten, M. (2000). Once upon a TIMSS: American and Japanese narrations of the Third International Mathematics and Science Study. *Education and Society*, 18(3), 45-76. <http://dx.doi.org/10.7459/es/18.3.05>
- Tienken, C. (2010). Common Core State Standards: I Wonder? *Kappa Delta Pi Record*, 47(1), 14-17. <http://dx.doi.org/10.1080/00228958.2010.10516554>
- Turque, B. (2011). *D.C.'s Bruce-Monroe school faces challenges as it tries Singapore math method*. Washington Post (Local section, June 6). Retrieved from http://www.washingtonpost.com/local/education/dcs-bruce-monroe-school-faces-challenges-as-it-tries-singapore-math-method/2011/06/01/AGuiHZKH_story.html
- U.S. Department of Education (2004). Achieving diversity: Race-neutral alternatives in American education. Retrieved from www.ed.gov/about/offices/list/ocr/edlite-raceneutralreport2.html
- Vygotsky, L. S. (1997). *The collected works of L. S. Vygotsky: Vol. 4. The history of the development of higher mental functions*. New York: Plenum Press.
- Wagner (2012). What should be learned from learning assessments? *Compare*, 42, 510-512.
- Walberg, H. J. (1984). Families as partners in educational productivity. *Phi Delta Kappan*, 65, 397-400.
- Wang, J. (2001). TIMSS primary and middle school data: Some technical concerns. *Educational Researcher*, 30(6), 17-21. <http://dx.doi.org/10.3102/0013189X030006017>
- White, M. (2012, May 25). Can U.S. schools adopt education practices of top-performing nations? Deseret News. Retrieved from <http://www.deseretnews.com/article/765578482/Can-US-schools-adopt-education-practices-of-top-performing-nations.html>
- Wise, S., & Demars, C. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, 10, 1-17. http://dx.doi.org/10.1207/s15326977ea1001_1
- Wise, S., & Demars, C. (2008, March). *Examinee non-effort and the validity of program assessment results*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.
- Woolard, J. C. (2013). Prelude to the Common Core: Internationally benchmarking a state's math standards. *Educational Policy*, 27, 615-644. <http://dx.doi.org/10.1177/0895904811429287>
- Wray, H. (1999). *Japanese and American education: Attitudes and practices*. West Port, CT: Bergin & Garvey.

- Wulf, W. (2005). The importance of foreign-born scientists and engineers to the security of the United States (Statement for a hearing by Committee on the Judiciary of U.S. House of Representatives on 15 September 2005). Retrieved from <http://www.aau.edu/WorkArea/DownloadAsset.aspx?id=6458>
- Yang, J. (2011). US students outperform Chinese in SATs. Retrieved from http://www.chinadaily.com.cn/china/2011-12/02/content_14204275.htm

Copyright Disclaimer

Copyright reserved by the author(s).

This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).