CHANCE FAVORS THE PREPARED MIND: *MATHEMATICS AND SCIENCE INDICATORS FOR COMPARING STATES AND NATIONS*[1]

Gary W. Phillips
Chief Scientist
American Institutes for Research
November 14, 2007

[1] Copies of this paper can be downloaded by searching www.air.org, and questions can be addressed to the author at gwphillips@air.org. Proper citation is as follows: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators for Comparing States and Nations*, American Institutes for Research: Washington, DC, 2007.

Table of Contents

List of Tables

List of Tables (Continued)

List of Tables (Continued)

List of Figures

List of Figures Continued

Executive Summary

- *This report provides international benchmarks to help states see how students are doing in math and science within an international context.*
- ***Good News**—Most states are performing as well or better than most foreign countries.*
- ***Bad News**—The highest achieving states within the United States are still significantly below the highest achieving countries.*

This paper describes state and international education indicators for mathematics and science using state data collected by the 2005 and 2007 National Assessment of Educational Progress (NAEP) and international data collected by the 2003 Trends in International Mathematics and Science Study (TIMSS) in grade 8.[2] Data from the two studies are expressed in the same metric through statistical linking (Phillips, 2007). By expressing both assessments in the same metric, states within the United States can use TIMSS results as international benchmarks to monitor progress over time. The overall findings at the national and state level were as follows.

National Level

- At the national level, several Asian countries generally outperform the United States in both mathematics and science, while many African and Middle Eastern Countries performed significantly below the United States. The United States was generally comparable to other English-speaking nations and European countries. The highest performing countries were also the same ones that grant the largest proportion of college degrees in science, technology, engineering, and mathematics (see figure 46).

- In mathematics, the means of five countries reached the Proficient level of achievement. These were Singapore, Hong Kong (SAR), Republic of Korea, Chinese Taipei, and Japan. Twenty-two countries were at the Basic level (including the United States), and 19 counties were Below Basic (see table 1).

- In science, only two countries had means that reached the proficient level of achievement. These were Singapore and Chinese Taipei. Twenty countries were at the Basic level (including the United States), and 24 countries were Below Basic (see table 3).

State Level

- At the state level, this report showed that although there is considerable variation in state performance, states are not as variable as nations. Even the highest achieving states within the United States were still significantly below the highest achieving countries,

[2] The NAEP results in this paper were obtained from publicly available data at www.nces.ed.gov . The state science results are from 2005, and the state mathematics results are from 2007 (the most recent NAEP assessments in each subject). The 2003 TIMSS results reported in this paper are based on publicly available data obtained from www.TIMSS.bc.edu, which were re-analyzed by Phillips (2007) and reported at www.air.org using NAEP achievement levels linked to the TIMSS scale.

and the lowest performing states were still significantly higher than the lowest achieving countries (see figures 1–53).

- In mathematics (in 2007 NAEP), no state average reached the Proficient level (although the Massachusetts mean is only one scaled score point away from reaching the Proficient level). Instead every state is performing at the Basic level with the exception of the District of Columbia, which is Below Basic (see table 2).

- In science (in 2005 NAEP), no state average reached the proficient level. The mean of thirty-five states (plus Department of Defense Education Activity) are at the Basic level. Nine state averages are at the Below Basic level (see table 4).

The paper argues that the United States needs to substantially increase the scientific and mathematical competency of the general adult population so that the voting citizenry can better understand and reach a consensus on policies that address many of the world's most pressing problems.

In addition we need larger numbers of people working in the scientific disciplines in order to better compete in a global economic environment. To achieve these goals, national and state policy makers need indicators of scientific and mathematical progress early in the educational pipeline. It is argued that the strategy of linking NAEP to TIMSS helps to provide this system of indicators.

Introduction

This paper shows how state-by-state results from the National Assessment of Educational Progress (NAEP) can be linked with nation-by-nation results from the Trends in International Mathematics and Science Study (TIMSS) to provide a comprehensive indicator system that would allow state-by-nation comparisons. Such a system of indicators is important to state and national policy makers because it goes beyond the traditional roles of NAEP and TIMSS. Historically, NAEP has allowed U.S. policy makers to compare and track the progress of states within the United States, while TIMSS has provided similar data for nations. This report places NAEP and TIMSS on the same scale, allowing states to compare themselves with nations. By doing so, states can monitor progress toward improved science and mathematics achievement while seeing how they stack up within an international context. This strategy is analogous to converting world currencies to dollars as an external benchmark for tracking local economic progress.

The paper first explores the broader context for the study by arguing that many intractable worldwide problems cannot be addressed in the United States until we reach a critical mass of science and mathematical literacy among the general population. Until the general population becomes aware of the science underlying these problems, they will not be able to establish public policy to address the solutions. In addition to needing more science and mathematics literacy among the general public, the United States needs more students preparing for careers in science, technology, engineering, and mathematics. To meet the demands of the future, a larger proportion of our workforce must have the problem solving and critical thinking skills to compete in a technologically sophisticated and global environment.

Monitoring progress toward reaching these goals needs to start early, while the cohort is still in the pipeline. Measuring students' knowledge of science and mathematics in the 8th grade is an ideal point in the pipeline to take the temperature of the progress. The 8th grade is probably the last year in which the student population broadly reflects the general population. After the 8th grade, public schools experience increasing dropout rates, and many countries direct students into vocational and academic tracks. Also, the end of middle school is a good time to find out how prepared students are to take further mathematics and science courses and possibly enter careers in science, technology, engineering, and mathematics.

The paper then discusses a brief history of attempts within the United States to establish state-by-state indicators of student performance. The paper argues that most attempts have been flawed. However, there is a way to use extant data from NAEP and TIMSS to provide a comprehensive indicator system with accurate and timely state-by-state data along with international benchmarks for states.

Finally, the paper introduces the concept of statistically linking NAEP and TIMSS. This allows TIMSS to be reported based on the NAEP achievement levels. By expressing NAEP and TIMSS in the same metric, states can see not only see how they compare with other states, but also with other countries.

Context for the Study

Low levels of scientific and mathematical literacy among the general public

To understand many of the world's most pressing problems, you must have a level of competency in science and mathematics. Furthermore, many of these problems can only be solved when the general citizenry has sufficient scientific and mathematical awareness to reach a consensus about what to do. Large societal issues such as global warming, deforestation, use of fossil fuels, population growth, ozone depletion, rising obesity rates and pandemic virus infections can only be addressed when enough people in the general population understand the science underlying these problems. Only then can they reach a national consensus about public policy.

According to the National Science Foundation (NSF, www.nsf.gov/statistics), the average U.S. citizen understands very little science. For example:

- Two-thirds do not understand DNA, "margin of error," the scientific process, and do not believe in evolution.
- Half do not know how long it takes the earth to go around the sun, and a quarter does not even know that the earth goes around the sun.
- Half think humans coexisted with dinosaurs and believe antibiotics kill viruses.

On the other hand, according to the NSF, the general public believes in a lot of pseudoscience.

- Eighty-eight percent believe in alternative medicine.
- Half believe in extrasensory perception and faith healing.
- Forty percent believe in haunted houses and demonic possession.
- A third believes in lucky numbers, ghosts, telepathy, clairvoyance, astrology, and that UFOs are aliens from space.
- A quarter believes in witches and that we can communicate with the dead.

The average citizen is also not very literate in mathematics. According to the National Center for Education Statistics (http://nces.ed.gov/naal/sample.asp):

- Seventy-eight percent cannot explain how to compute the interest paid on a loan.
- Seventy-one percent cannot calculate miles per gallon on a trip.
- Fifty-eight percent cannot calculate a 10% tip for a lunch bill.

The latest results of the National Assessment of Adult Literacy (NAAL) in 2003 revealed very low levels of quantitative literacy among American adults. Using performance standards developed by the National Academy of Sciences (Hauser et al, 2005) only 13% of adults were at the highest level of proficiency. Furthermore, there had been no change in this level of literacy from 2002 to 2003. An example of the mathematics skill required at the highest level of proficiency is, "can the person compute and compare the cost per ounce of a food item?" (Kutner et al, p. 3)

In a democracy, a critical mass of the general population needs to grasp complex concepts in sufficient detail to make informed societal decisions. Furthermore, with the growth of globalization, the pressure of international competition, and the impending retirement of millions of baby boomers, state and national policy makers need to worry about the quality of the next generation of students who are currently in the educational pipeline.

Lack of preparation of students for careers in science, technology, engineering, and mathematics (STEM)

In addition to needing more science and mathematics literacy among the general public, the United States needs more students preparing for careers in science, technology, engineering, and mathematics (STEM). The future workforce must have substantially more innovation, problem solving, and critical thinking skills to compete in a technologically sophisticated and global environment. The concern is that there are not enough students in the educational pipeline who are prepared to work in these areas. According a recent General Accounting Office (GAO) report, postsecondary education enrollment has increased over the past decade, but the percentage of students obtaining degrees in STEM fields has declined (GAO, 2006). Only 16% of all postsecondary degrees in the United States are STEM-related (NCES, 2005), and many of these are awarded to foreign students. Furthermore, "a significant number of university faculty in the scientific disciplines are foreign, and foreign doctorates are employed in large numbers by industry" (CRS, 2006, p. 14).

In fact, the United States has one of the lowest proportions of STEM *first university* degrees (16.8%) awarded among the countries surveyed by the NSF (2006). The race to prepare students in the pipeline for the future is clearly being won by our Asian economic competitors, with China at 52.1%, Japan at 64%, and South Korea at 40.6%. Furthermore, even though the United States has a very high rate of postsecondary education attainment, it still ranks below Japan and China in the absolute number of STEM degrees awarded (CRS, 2006, p. 17).

Preparing the 50 million students enrolled in our 97,000 public schools is done at an annual expense of 500 billion dollars to the American taxpayer. How do we know we are getting our money's worth? Are we getting results? Is there an indicator of success? For example, how well does the mathematical and scientific competency of our states and the nation stack up against our major economic competitors, such as members of the Group of Eight (referred to as the G8—Canada, France, Germany, Italy, Japan, the Russian Federation, the United Kingdom, and the United States). State and national education policy makers need international benchmarks against which state and national performance can be gauged.

Indicators

There are many types of indicators and benchmarks that policy makers need to understand different educational systems and to identify reform strategies to improve student achievement within the United States. For example, there is a need for high-quality information on indicators related to expenditures, enrollment, attainment, quality of the teacher workforce, opportunity to learn, and other indicators of access and equity. By far, the most important indicators that are needed are outcome measures that relate to the success of educational systems. *This type of "outcome" indicator is the focus of this report.*

What is an indicator? The word "indicator" comes from the Latin verb *indicare*, which means “to disclose” or “to point out.” An indicator is like a sign post. It helps you understand where you are, whether you are going in the right direction, and how far you are from where you want to be. When we travel, we use signposts to help stay on course. They do not provide as much information as a map, but they help alert you to problems before you get lost. A signpost will help you recognize the direction you need to go to get back on course. Similarly, an indicator of state educational success would help policy makers determine whether they are going in the right direction and how far they are from where they want to be.

What are the characteristics of a *good* national and state-by-state outcome indicator?

- *First*, we probably want the indicator to be a single number (so it is easy to understand and remember) and comparable across units being compared. Normally it is a statistic, an index, a weighted average, or a composite of several variables. Some examples of this outside the education realm would be using the consumer price index (CPI) as a measure of the price of goods and services and a monitor for inflation as well as using the gross domestic product (GDP) as a measure of the size of a nation’s economy.

- *Second*, we want the indicator to be accurate so there is no question about its reliability and validity. Since statistical accuracy is one of the primary roles of statistical agencies within the United States, it is probably a good idea to rely on numbers obtained from data collected through surveys by those agencies.

- *Third*, a good indicator has some causal connection to the phenomenon of which the number is an indicator. Consequently, the indicator is a sign, symptom, or summary measure of a phenomenon.

- *Fourth*, the indicator should show direction. Is the phenomenon going up or going down, and are we making progress or falling behind?

- *Fifth*, the indicator should be something that is empirically external to the user of the index (is not influenced by the user’s actions). In other words, it should be an index that is not corruptible by the actions of the people affected by it.

- *Sixth*, a national and state-by-state education indicator should have international benchmarks. A nation and a state should be able to see how they stack up against educational systems around the world.

It should be noted that there are several things that good national and state-by-state educational indicators *cannot* do. Policy makers should not confuse useful indicators with useful *goals.* Indicators can help monitor progress toward useful goals but cannot make the goals useful. Furthermore, indicators of program effects are not the same thing as *evaluations* of program effects (that requires data designed and collected for that purpose). Finally, indicators are not a substitute for educational *research* because they only provide correlation information between variables and do not provide information about causal connections (e.g., that might require designs such as randomized trials).

Currently, there are a large number of organizations reporting sets of education indicators. For example, there are many *national* and *international indicator* systems available. Among them are those used by the National Center for Education Statistics (NCES), at www.nces.ed.gov; the Organization for Economic Co-operation and Development (OECD) International Indicators of Education Systems (INES) Project, at www.oecd.org: the United Nations Educational, Scientific and Cultural Organization (UNESCO) World Education Indicators (WEI), at www.uis.unesco.org; and the World Bank, at www.worldbank.org.

Similarly, there are a large number of agencies and organizations reporting state-by-state indicators. Among the *state-by-state indicator* systems are those provided by the National Assessment of Education Progress, at www.nces.ed.gov/nationsreportcard, the Council of Chief State Officers (CCSSO) State Education Indicators, at www.ccsso.org; Education Week, at www.edweek.org; the U.S. Chamber of Commerce, at www.uschamber.com; and the National Center for Public Policy & Higher Education's Measuring Up, at www.highereducation.org.

Brief History of State and National Education Indicators

The realization that the United States needed better statistical indicators of educational performance gradually emerged as part of the search for "social indicators" in the late 1960s and early 1970s. This effort was institutionalized in 1974 when Congress authorized the creation of the annual *Condition of Education* report. However, the special focus on state-by-state education indicators was likely "jump-started" by *A Nation at Risk* in 1983.

A Nation at Risk (1983)

In 1983, the U.S. Department of Education's National Commission on Excellence in Education (a blue-ribbon commission appointed under the Reagan administration) published the report, *A Nation at Risk: The Imperative for Education Reform.* The disturbing language in this document is often credited with being the pebble that started the waves of national education reform we still see today. The language was direct and dire.

"*Our Nation is at risk. Our once unchallenged preeminence in commerce, industry, science, and technological innovation is being overtaken by competitors throughout the world. …We report to the American people that while we can take justifiable pride in what our schools and colleges have historically accomplished and contributed to the United States and the well-being of its people, the educational foundations of our society are presently being eroded by a rising tide of mediocrity that threatens our very future as a Nation and a people. What was unimaginable a generation ago has begun to occur—others are matching and surpassing our educational attainments.*

If an unfriendly foreign power had attempted to impose on America the mediocre educational performance that exists today, we might well have viewed it as an act of war. As it stands, we have allowed this to happen to ourselves. …We have, in effect, been committing an act of unthinking, unilateral educational disarmament." (*A Nation at Risk: The Imperative for Education Reform,* April 1983).

The report was a huge media success and helped mobilize public support to rally around education reform.

"*A Nation at Risk and the other education reports of the early 1980s helped launch the first wave of educational reforms that focused on expanding high school graduation requirements, establishing minimum competency tests, and issuing merit pay for teachers*." (Vinovskis, 1999).

Following the publication of *A Nation at Risk,* it gradually became clear to governors and other policy makers that improving their educational systems would not be possible without state-by-state data that were comparable, reliable, and timely. How else could a governor prove to the public that the increased investment in reform led to improved student achievement? Unfortunately, there was no readily available set of indicators that did not give a misleading impression of state-by-state educational performance.

Wall Chart (1984–1989): SAT and ACT as state-by-state indicators

The first attempt to piece together a set of state-by-state outcome indicators was the 1984 publication of the "Wall Chart" by the U.S. Department of Education.

"In 1984 the wall chart of State Education Statistics broke the historic silence on reporting state-by-state comparisons of student performance. Prior to its release chief state school officers and the education establishment had been protected from disclosure of poor performance by the states in education. The wall chart, by laying out the facts in straightforward detail, exposed our national shortcomings in education and focused our attention on the states where much of the education policymaking takes place." (Ginsburg, Noell, and Plisko, 1988).

The Wall Chart used average state aggregates of SAT and ACT scores. The Wall Chart was used even though it was widely criticized because it only measured the self-selected college-bound population. The larger the percentage of the population taking the SAT or ACT tests, the lower the state's ranking on the Wall Chart. The states with the least number of students heading for college tended to have the highest ranking. In fact, the 1986 correlation between the SAT and the proportion of college-bound students were –0.86 (College Board, 1986). The fact that it was a biased indicator due to self-selection did not deter the department from using the system for six years under two secretaries of education, Terrell H. Bell and William J. Bennett.

"*Some analysts see state-by-state comparisons as filling a void in our statistical knowledge, enabling states and their residents to gauge for the first time the quality of their education. Others see this information as statistically flawed and providing little guidance to improve the system; worse yet, they say, the measures may mislead, sending reform efforts off in the wrong direction. We believe that the publication of the wall chart, with its acknowledged flaws, has helped validate state-by-state comparisons as a means of holding state and local school systems accountable for education.*" (Ginsburg, Noell, and Plisko, 1988).

The Wall Chart created considerable debate and helped the country focus attention on the fact that there were no good state-by-state measures of educational achievement.

Lake Wobegon Report (1987): NRTs as state-by-state indicators

In 1987, a West Virginia physician produced a report of the results of a survey where he had found that on norm-referenced tests (NRTs), all 50 states were above the national average (Cannell, 1987, 1988). This sparked much interest in Washington because it was hoped that NRTs might overcome some of the problems of the SAT and ACT as indicators of state-by-state performance. Since they had national norms, were administered under standardized conditions, and given in many states to a census of students, it was hoped the self-selection issues of the SAT and ACT could be overcome. The report made the front page of both *The Washington Post* (Feinberg, 1988) and *The New York Times* (Fiske, 1988). Ultimately, a special issue of *Educational Measurement: Issues and Practice* was devoted to the topic (Vol. 7(2), Summer 1988), and it became the topic of countless educational testing conferences. In 1988, the U.S. Department of Education sponsored a meeting of the major NRT publishers. At the meeting there were many criticisms of the methodology and inferences from the report. However, there was,

"Unanimous agreement that the primary finding (that all fifty states were above the national average in the elementary grades) was correct." (Phillips, 1990).

The major explanation provided was that some norms used by states were outdated and, over time, teachers became familiar with the test items and taught to the test. Regardless of the reason, it became clear that comparing states based on NRTs was fundamentally flawed.

No Child Left Behind (2001–present): CRTs as state-by-state indicators

On January 8, 2002, the No Child Left Behind Act of 2001 (NCLB) was signed into law. The legislation required states to develop content standards, achievement standards, and achievement tests in reading and mathematics for grades 3–8 and one grade in high school. In practice, each state develops its own content standards, its own achievement standards, and its own criterion-referenced test (CRT), so there is no comparability across states. It is obvious that such state-developed CRT results cannot be used as indicators for state-by-state comparisons. For example, in 2005, Georgia, Oklahoma, and South Carolina each had 26% of their 4th-grade students classified as Proficient or above on the state NAEP reading assessment. However, on the state CRT, Georgia had 85%, Oklahoma had 83%, and South Carolina had 35% (Vu, 2007). This leads to such statements as *"Johnny can't read ... in South Carolina. But if his folks move to Texas, he'll be reading up a storm"* (Petersen & Hess, 2005). Under NCLB, states can develop their own tests and set different standards, but call them by the same name. This is a kind of "jabberwocky" that obfuscates accountability at the national level and renders state-by-state comparisons virtually uninterpretable.

Not only are the state-by-state comparisons with CRTs uninterpretable (because of variation in state performance standards) but they are also misleading. Because NCLB requires states to make adequate yearly progress (AYP) incrementally increasing to 100% proficiency in 2013-2014, states are motivated to set low standards. This was demonstrated by a recent report that mapped 2005 state-developed proficiency standards on to the National Assessment of Educational Progress (NAEP) scale (NCES, 2007). For example, the report correlated the 2005 8th-grade math performance standard on the state test with the NAEP score that was equivalent to

the state standard in 36 states. The report found a high negative correlation of –.83 between the proportion meeting the state standard and the state standard projected on to the NAEP scale. *This means high performance on the state test is associated with low standards on NAEP*. This report disentangled the differences in the stringency of the local state standard from the differences in the distributions of skills of the state population of students. It was shown that the reason states have substantially different proportions of proficient students is largely due to differences in standards rather than difference in student performance.

Policy makers need state-by-state data to guide them in efforts to improve learning and monitor accountability. It is clear that there is something terribly wrong with America's extant, piecemeal, locally controlled, state education data system. How can policy makers use norm-referenced tests to compare states if all the states are above the national average? How can they use state developed criterion-referenced tests if the highest levels of proficiency are reported in the states with the lowest standards? How do we know if we are making progress? How do we know if one state is performing better than another? It turns out that state criterion-referenced tests, lead to the same epistemological conundrum as their cousin, the national norm-referenced test. Without an independent, reliable, comparable, external referent, state policy makers will never get out of Lake Wobegon.

Using state-developed CRTs as state-by-state indicators clearly violates the first and fifth characteristic of a good state-by-state indicator, as previously mentioned. That is, the indicator should be something that is comparable across states and be empirically external to the user of the index (i.e., the state). State CRTs are important monitors of within-state progress, but they should not be used to compare states.

NAEP State Assessment (1990–present): NAEP as state-by-state indicator

In May 1986, Secretary of Education William Bennett created a 22-member panel to review the NAEP to see if it could be improved to monitor educational progress. The panel was headed by Tennessee Governor Lamar Alexander (who was also the chair of the National Governors' Association) and H. Thomas James (former president of the Spencer Foundation). The panel is often referred to as the Alexander/James Study Group. In January 1987, the panel released its report, often referred to as the Alexander/James report.

"The single most important change recommended by the Study Group is that the assessment collect representative data on achievement in each of the fifty states and the District of *Columbia. Today state and local school administrators are encountering a rising public demand for thorough information on the quality of their schools, allowing comparison with data from other states and districts and with their own historical records. Responding to calls for greater accountability for substantive school improvements, state officials have increasingly turned to the national assessment for assistance."*
(Alexander/James Study Group, 1987, p. 11–12)

The Alexander/James report became the blueprint for the reorganization of NAEP within the reauthorization of the Elementary and Secondary Education Act of 1965 (P.L. 89-10). The final legislation, the Augustus F. Hawkins-Robert T. Stafford Elementary and Secondary School Improvement Amendments of 1988 (P.L. 100-297), created limited state-level NAEP testing on

a voluntary and *trial* basis in mathematics and reading for those states choosing to participate. The first trial state assessment was conducted in 1990 in 8th-grade mathematics and released on June 6, 1991, at the National Press Club in Washington, DC (Mullis, Dossey, Owen, & Phillips, 1991). The Press Club was packed to capacity; the release was covered by every major newspaper in the country and was on the front page of many of them.

In future assessments, more grades and subjects were added and more states participated, and in 1996, the authorizing legislation no longer treated the state assessments as a trial. In 2001, with the reauthorization of the Elementary and Secondary Education Act (referred to as "No Child Left Behind"), in order to receive Title I funding, states were required to participate every two years in state NAEP in reading and mathematics at grades 4 and 8. This legislative act pretty much guaranteed that all states would participate in NAEP.

State NAEP is the ideal national and state-by-state indicator of educational progress. Because state NAEP is legislatively mandated and funded, developed by a national consensus process, overseen by an independent policy board (the National Assessment Governing Board—NAGB), and administered by an independent statistical agency (NCES), it represents the CPI of education. It just needs one more ingredient—an external international benchmark.

International Assessments: TIMSS as nation-by-nation indicator

The first international assessments were conducted by the International Association for the Evaluation of Educational Achievement (IEA). The IEA is currently located in the Netherlands and has been the main source of international data over the past 50 years. For at least the first 30 years, the IEA studies were episodic. The irregular intervals of the studies made them useful for researchers but not useful for governments who needed regular, reliable, and timely data. Because governments were not too involved in these studies, the IEA studies were poorly funded and therefore could take up to a decade to collect, analyze, and report the results. Beginning in 1989, NCES decided it needed international data on a regular basis. Also, the needs of governments were broader in scope than what the IEA studies provided. Rather than focusing on in-depth analyses of within-country educational achievement, NCES wanted data that would facilitate cross-country comparisons and be linkable to NAEP. To accomplish this, NCES funded the first study of the International Assessment of Educational Progress (IAEP), which was conducted in February 1988. The study used the NAEP content standards and was administered in five countries and four Canadian provinces. In 1991, the IAEP was expanded to 20 countries.

Shortly after the release of the second IAEP results, the IEA submitted to NCES a proposal to conduct a third IEA mathematics study. NCES felt the study was too much like the old IEA studies (representing a lot of in-depth, time-consuming research) and needed to be more like the IAEP studies (representing a broad indicator type of information). NCES laid out the design parameters of the next international study it wanted to fund. It should be in grades 4 and 8, cover both mathematics and science, use content standards based on a broad international consensus, be on a 4-year cycle, and be linkable to NAEP. This design was discussed and accepted at a meeting of the Board on Testing and Assessment (BOTA) at the National Academy of Sciences (NAS). In attendance at the BOTA meeting was the U.S. national representative to the IEA. Within several days, the IEA resubmitted a proposal to NCES titled the *Third International*

Mathematics and Science Study (TIMSS).[3] The first TIMSS was conducted in 1995 (in 45 countries), with follow-up studies conducted in 1999, 2003, and 2007.

This report has argued that NAEP state assessments have all the characteristics of an excellent indicator for state-by-state comparisons. Similar arguments could be made for TIMSS providing a good indicator of nation-to-nation comparisons. Can the two be combined so that we can compare states to states, nations to nations, and states to nations? This is where a statistical linking study comes in.

NAEP Linked to TIMSS: State-by-nation indicators

What is most relevant in this brief chronology of TIMSS is that it was purposely designed to be linkable to NAEP. It is by design, and not by accident, that both TIMSS and NAEP are conducted in the same grades, cover similar content standards, use matrix sampling of cognitive items, use similar background items to address policy questions, use similar nationally representative sampling techniques, use similar scaling models (item-response theory), and use similar analysis models (plausible values).

The use of statistical linking as a way to connect NAEP to external assessments was foreshadowed by the Alexander/James Study Group. Following the recommendation for assessments at the state level, the report recommended that NAEP establish *linkages* with other local, state, and international assessments.

"*Recent developments in test theory and measurement technology now make it possible to compare scores from different assessment instruments, thus broadening the scope of comparisons that can be made. We recommend that the national assessment devise a linkage system relating local and state testing and assessment programs to the national assessment…Recent years have also witnessed an increasing interest in the use of national assessment data for international comparisons of student performance.*" (Alexander/James Study Group, 1987, p. 12–13)

Conceptually, linking two assessments simply means the two are connected in such a way that there is a cross-walk between them (e.g., a cross-walk between NAEP and TIMSS) that allows you to compare their results. Linking is a statistical procedure that allows you to express the results of one test (e.g., TIMSS) in terms of the metric of another (e.g., NAEP). Once the link is established, results of each assessment can be compared (e.g., the results of states on NAEP can be compared to the results of nations on TIMSS). In the physical sciences, this is similar to expressing Fahrenheit in terms of Celsius. The cross-walk is the equation $F^o = 32 + 1.8\left(C^o\right)$. The cross-walk between NAEP and TIMSS is more complicated, and of course has considerably more error, than the cross-walk between temperature metrics. The determination of this cross-walk, and error, are the primary outcomes of statistical linking studies.[4]

[3] The definition of the acronym TIMSS was subsequently changed to Trends in International Mathematics and Science Study.

[4] For details of the linking procedure, see the technical Appendix A and Phillips (2007).

Although there have been several previous studies in which NAEP has been statistically linked to international assessments, there has only been one prior study that used the link to compare NAEP state achievement level results with international results. This was the Pashley and Phillips (1993) study which linked the 1991 IAEP (age 13) and 1992 NAEP (grade 8) in mathematics. The study was used to estimate how other countries who took the IAEP stacked up against the NAEP achievement levels. In the paper, both the 15 countries in the 1991 IAEP and all the states that participated in the 1990 and 1992 state NAEP were analyzed in terms of their performance on the NAEP achievement levels.

The present study uses the results of a recently released report by this author (Phillips, 2007). The Phillips study linked the NAEP achievement levels to the TIMSS scale in 8th-grade mathematics and science using data from the 2000 NAEP and the 1999 TIMSS.

The definition of the 8th-grade NAEP proficient achievement levels *in mathematics* is provided in the NAEP 2000 mathematics report (Braswell et al. 2001, p. 11). The first sentence of the definitions is referred to as the policy definition of the achievement level.

> ***Basic*** level denotes partial mastery of the knowledge and skills that are fundamental for proficient work at a given grade. Eighth-grade students performing at the *Basic* level should exhibit evidence of conceptual and procedural understanding in the five NAEP content strands (number sense, properties, and operations; measurement; geometry and spatial sense; data analysis, statistics, and probability; and algebra and functions). This level of performance signifies an understanding of arithmetic operations—including estimation—on whole numbers, decimals, fractions, and percents.
>
> ***Proficient*** level represents solid academic performance. Students reaching this level demonstrate competency over challenging subject matter. Eighth-grade students performing at the *Proficient* level should apply mathematical concepts and procedures consistently to complex problems in the five NAEP content strands (number sense, properties, and operations; measurement; geometry and spatial sense; data analysis, statistics, and probability; and algebra and functions).
>
> ***Advanced*** level signifies superior performance at a given grade. Eighth-grade students performing at the *Advanced* level should be able to reach beyond the recognition, identification, and application of mathematical rules in order to generalize and synthesize concepts and principles in the five NAEP content strands (number sense, properties, and operations; measurement; geometry and spatial sense; data analysis, statistics, and probability; and algebra and functions).

The definition of the 8th-grade NAEP proficient achievement level in *science* is provided in the NAEP 2000 science report (O'Sullivan et al. 2003, p. 12).

> ***Basic*** level denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade. Students performing at the *Basic* level demonstrate some of the knowledge and reasoning required for understanding of the Earth, physical, and life sciences at a level appropriate to grade 8. For example, they can carry out investigations and obtain information from graphs, diagrams, and tables. In addition, they demonstrate some understanding of concepts relating to the solar system and relative motion. Students at this level also have a beginning understanding of cause-and-effect relationships.

Proficient level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter. Students performing at the *Proficient* level demonstrate much of the knowledge and many of the reasoning abilities essential for understanding of the Earth, physical, and life sciences at a level appropriate to grade 8. For example, students can interpret graphic information, design simple investigations, and explain such scientific concepts as energy transfer. Students at this level also show an awareness of environmental issues, especially those addressing energy and pollution.

Advanced level signifies superior performance. Students performing at the *Advanced* level demonstrate a solid understanding of the Earth, physical, and life sciences as well as the abilities required to apply their understanding in practical situations at a level appropriate to grade 8. For example, students can perform and critique the design of investigations, relate scientific concepts to each other, explain their reasoning, and discuss the impact of human activities on the environment.

Results

The results of this report for grade 8, mathematics and science, are contained in the 53 figures[5] that follow as well as tables 1–4.

In each figure, the percent at and above Proficient from the NAEP was obtained from the publicly available data at www.nces.ed.gov. The international results are from Tables 23 and 24 in Appendix A. Figures 1–53 display state-by-nation indicators of mathematics and science performance. These figures provide the international benchmarks states need in order to see how they stack up against international competitors.

The figures are arranged in alphabetical order by state. In each figure, state results from the 2007 state NAEP in mathematics and 2005 state NAEP in science (the most recent state NAEP assessments in each subject) are compared to all the nations in the 2003 TIMSS (the exception is Figure 46, which shows the United States NAEP compared to each nation). These state-by-nation comparisons are made possible by the NAEP-TIMSS linking study (Phillips, 2007).

U.S. National Results

The results for the United States are contained in Figure 46. The graphs indicate which nations are statistically above, similar to, and below the United States.[6] This is indicated by the taller black bars on the left, white bars in the middle, and shorter black bars on the right, respectively.

[5] I would like to thank Futoshi Yumoto, Jeff Foarde and James Phillips for assistance with the graphs.

[6] To be consistent with NAEP, this paper uses adjustments for multiple comparisons for statistical significance testing. Please see technical Appendix B for details.

International Benchmarks for the United States in Mathematics

For mathematics, we see that six nations have significantly more students who meet the Proficient standard. These are

1. Singapore,
2. Hong Kong (SAR),
3. the Republic of Korea,
4. Chinese Taipei,
5. Japan, and
6. Belgium (Flemish).

There are 8 nations with mathematics performance similar to the United States. These include

1. Netherlands,
2. Hungary,
3. Estonia,
4. Slovak Republic,
5. Australia,
6. Russian Federation,
7. Malaysia, and
8. Latvia.

There are 31 countries that are significantly below the United States in their percentages of proficient mathematics students. These are

1. Lithuania,
2. Israel,
3. England,
4. Scotland
5. New Zealand,
6. Sweden,
7. Serbia,
8. Slovenia,
9. Romania,
10. Armenia,
11. Italy,
12. Bulgaria,
13. Republic of Moldova,
14. Cyprus,
15. Norway,
16. Republic of Macedonia,
17. Jordan,
18. Egypt,
19. Indonesia,
20. Palestinian National Authority,

21. Lebanon,
22. Islamic Republic of Iran,
23. Chile,
24. Bahrain,
25. Philippines,
26. Tunisia,
27. Morocco,
28. Botswana,
29. South Africa,
30. Saudi Arabia, and
31. Ghana.

Many of these nations have proficient levels in the single digits, and four nations have no one that could be statistically surveyed as functioning at the Proficient level. These nations are Botswana, South Africa, Saudi Arabia, and Ghana.

International Benchmarks for the United States in Science

Also in Figure 46 are the overall national results in science which are similar to mathematics. For science, eight nations perform significantly better than the United States. These are

1. Singapore,
2. Chinese Taipei,
3. Republic of Korea,
4. Hong Kong (SAR),
5. Japan,
6. Estonia,
7. England, and
8. Hungary.

Ten countries have science performance similar to the United States. These are

1. Netherlands,
2. Australia,
3. Sweden,
4. New Zealand,
5. Slovak Republic,
6. Lithuania,
7. Slovenia,
8. Russian Federation,
9. Scotland,
10. Belgium (Flemish),

Finally, 27 countries perform significantly below the United States in science. These are

1. Latvia,
2. Malaysia,
3. Israel.
4. Bulgaria,
5. Italy,
6. Jordan,
7. Norway,
8. Romania,
9. Serbia,
10. Republic of Macedonia,
11. Republic of Moldova,
12. Armenia,
13. Egypt,
14. Palestinian National Authority,
15. Islamic Republic of Iran,
16. Cyprus,
17. Bahrain,
18. Chile,
19. Indonesia,
20. Philippines,
21. Lebanon,
22. Saudi Arabia,
23. Botswana,
24. South Africa,
25. Morocco,
26. Ghana, and
27. Tunisia.

The low performance of many of these nations is similar to their performance in mathematics, with many of these nations having Proficient levels in the single digits, and two nations having no one that could be statistically surveyed as functioning at the Proficient level. These nations are Ghana and Tunisia.

State-by-Nation Results

Figure 1 will be used to illustrate state results. In Figure 1, we have a comparison between Alabama in the 2005 (science) and 2007 (mathematics) state NAEP as well as between each nation in the 2003 TIMSS.

There are two graphs in Figure 1. The first displays the results for grade 8 mathematics in Alabama compared to each nation. The second graph displays similar data for science. For each nation, the graph displays the percentage of students estimated to be at and above Proficient. The nations in each graph have been rank-ordered, with the highest achieving nations on the left and

the lowest performing countries on the right. Embedded within the graph is the percent at and above Proficient for Alabama.

International Benchmarks for Alabama in Mathematics

We see that there are 17 countries performing statistically better in mathematics than Alabama (indicated by the taller black bars to the left of Alabama). They are

1. Singapore,
2. Hong Kong,
3. Republic of Korea,
4. Chinese Taipei,
5. Japan,
6. Belgium (Flemish),
7. Netherlands,
8. Hungary,
9. Estonia.
10. Slovak Republic,
11. Australia,
12. Russian Federation,
13. Malaysia,
14. United States TIMSS,
15. Latvia,
16. Lithuania, and
17. Israel.

There are 10 countries that have mathematics performance statistically similar to Alabama (indicated by the white bars surrounding Alabama). These are

1. England,
2. Scotland,
3. New Zealand,
4. Sweden,
5. Serbia,
6. Slovenia,
7. Romania,
8. Armenia,
9. Italy, and
10. Bulgaria.

It should be noted that the mathematics results for Alabama in Figure 1 are 2007 state-by-state NAEP results from the publicly available data at www.nces.ed.gov. The national results in Figure 1 (including the one labeled “United States TIMSS”) refer to the U. S. performance on the 2003 TIMSS, as reported by Phillips (2007). The 2007 U.S. NAEP average for the percent at and above Proficient for mathematics for public school students is 31%. Significance testing for

Alabama between the state NAEP and the national NAEP can be conducted by using the NAEP Data Explorer at www.nces.ed.gov.

There are 19 countries that perform significantly below Alabama in mathematics (indicated by the shorter black bars to the right of Alabama). These are

1. Republic of Moldova
2. Cyprus,
3. Norway,
4. Republic of Macedonia,
5. Jordan,
6. Egypt,
7. Indonesia,
8. Palestinian National Authority,
9. Lebanon,
10. Islamic Republic of Iran,
11. Chile,
12. Bahrain,
13. Philippines,
14. Tunisia,
15. Morocco,
16. Botswana,
17. South Africa,
18. Saudi Arabia, and
19. Ghana.

International Benchmarks for Alabama in Science

The graph for science can be interpreted in the same way for Alabama. In science, there are 12 nations achieving significantly higher than Alabama. They are

1. Singapore,
2. Chinese Taipei,
3. Republic of Korea,
4. Hong Kong, SAR,
5. Japan,
6. Estonia,
7. England,
8. Hungary,
9. United States TIMSS,
10. Netherlands,
11. Australia, and
12. Sweden.

The science results for Alabama in Figure 1 are 2005 state-by-state NAEP results from the publicly available data at www.nces.ed.gov. The national results in Figure 1 (including the one

labeled "United States TIMSS") refer to the U.S. performance on the 2003 TIMSS, as reported by Phillips (2007). The 2005 United States NAEP average for the percent at and above Proficient for science for public school students is 27%. Significance testing for Alabama between the state NAEP and the national NAEP can be conducted by using the NAEP Data Explorer at www.nces.ed.gov.

There are 14 nations which have performance in science similar to Alabama. They are

1. New Zealand,
2. Slovak Republic,
3. Lithuania,
4. Slovenia,
5. Russian Federation,
6. Scotland,
7. Belgium (Flemish),
8. Latvia,
9. Malaysia,
10. Israel,
11. Bulgaria,
12. Italy,
13. Jordan, and
14. Norway.

Finally, there are 20 nations performing significantly below Alabama in science.[7]

1. Romania,
2. Serbia,
3. Republic of Macedonia,
4. Republic of Moldova,
5. Armenia,
6. Egypt,
7. Palestinian National Authority,
8. Islamic Republic of Iran,
9. Cyprus,
10. Bahrain,
11. Chile,
12. Indonesia,
13. Philippines,
14. Lebanon,
15. Saudi Arabia,

[7] In some graphs, a nation that is ranked farther away from the state is not significantly different from the state, whereas a nation ranked closer to the state is deemed significantly different from the state. For example, for mathematics in figure 41, New Zealand (with 21% at and above Proficient) is not significantly below Rhode Island (28%), but Scotland (with 22% at and above Proficient) is significantly below Rhode Island. This is because the standard error for The New Zealand is larger than that of Scotland. This results in Scotland being significantly below Rhode Island, whereas the New Zealand is not.

16. Botswana,
17. South Africa,
18. Morocco,
19. Ghana, and
20. Tunisia.

The analysis for Alabama can be repeated for every state. Each state tells a different story. Which countries are important as international benchmarks for one state may be different for another state.

One general conclusion from the data is that the majority of states are performing as well or better than a large portion of the foreign countries surveyed. This is true in mathematics as well as science.

Another overall pattern among the states is that all states are performing below our Asian economic competitors. This is true of even our highest performing states. In other words, our highest performing states are significantly below the highest performing foreign countries. Instead, most states are comparable in performance to most European and English-speaking nations. Our lowest achieving states, however, generally still outperform the extremely low single-digit performance of most Middle Eastern and African nations.

Criterion-Referenced Interpretations

All of the above results are essentially *norm-referenced* interpretations of national and state performance. Comparing the percent Proficient between states and nations is informative and helps contextualize state-by-state comparisons with international benchmarks. But it does not tell us how well states and nations are doing compared to an absolute standard. For example, the national percent Proficient for the United States 2007 mathematics was 31% and in 2005 science was 27%. How good is that? Is that good enough? One *criterion-referenced* strategy for answering these questions is to examine the achievement level associated with the state or national average. If the state or national average has reached the Proficient level - that means the average (or typical) student is Proficient. Here we are answering the question "*is the average student in a state or nation Proficient in mathematics and science or are they achieving at a Basic or Below Basic level?*" The criterion-referenced description of what it means for the average student to be Proficient, Basic, or Below Basic can be obtained from the definitions of these achievement levels above (or see NAEP reports for more extensive descriptions). This type of information is presented in tables 1–4.

Table 1 provides the achievement levels associated with the mean for each nation in the 2003 TIMSS in mathematics. We see that the mean of five countries reached the Proficient level of achievement. These were Singapore, Hong Kong (SAR), Republic of Korea, Chinese Taipei, and Japan. Twenty-two countries were are at the Basic level (including the United States) and 19 counties were Below Basic.

Table 2 shows that *in mathematics in 2007 NAEP, no state average reached the Proficient level* (although the Massachusetts mean is only one scaled score point away from reaching the

Proficient level). *Furthermore,* e*very state is performing at the Basic level with the exception of the District of Columbia which is Below Basic.*

Table 3 reports on the achievement levels associated with the mean for each nation in the 2003 TIMSS in science. The mean of only two countries reached the Proficient level of achievement. These were Singapore and Chinese Taipei. Twenty countries were at the Basic level (including the United States), and 24 countries were Below Basic.

Table 4 presents similar information for states in the 2005 NAEP for science. *In science, no state mean has reached the Proficient level*. *The mean of thirty-five states (plus DoDEA) are at the Basic level*. *Nine state averages are at the Below Basic level*. Overall the performance in science is lower than in mathematics. The reader might be tempted to conclude that nations and states are not learning as much science as they are mathematics, but this may not be true. Phillips (2007, p. 13) provides evidence that the NAEP science achievement level is set higher than the mathematics standard.

There is an important additional finding from the criterion-referenced interpretation. Figures 1–53 above generally show that states are in the middle of the pack in comparison to foreign national performance. In other words, we are not excelling but we are not behind either. The criterion-referenced perspective shows that that the middle of the pack is not a very satisfactory place to be because it represents a Basic and Below Basic level of achievement. It falls short of the Proficient standard that is our goal.

Discussion and Conclusions

This paper demonstrates that it is possible to piece together (through a statistical linking strategy) results from NAEP and TIMSS to create a comprehensive state, national, and international index of student performance in mathematics and science. The index is the percent at and above Proficient, as defined by the NAEP achievement levels. By statistically linking NAEP to TIMSS, these same achievement levels can be located on the TIMSS scale, permitting the index to be calculated across all the nations that participate in TIMSS. The index meets all six criteria above for a good indicator.

(1) Each state has a single number (one for mathematics and one for science) that is easy to understand (percent at and above Proficient) and that serves as an overall index for the state.
(2) The indicator is funded and monitored by NCES, a statistical agency dedicated to maintaining the reliability and validity of the data.
(3) The indicator is a direct measure of what students are learning in the 8^{th} grade in mathematics and science. The contents of both the NAEP and TIMSS are determined through a national consensus process. Consequently, there is a broad consensus that the indicator is causally connected to the phenomena of interest.
(4) The indicator reflects progress over time. In fact, measuring progress is the fundamental mandate of both NAEP and TIMSS.
(5) The indicator is external to the states and nations that participate in the survey. The states and nations cannot select the samples, alter the test administration, or select the test items

in such a way as to give them an advantage. Consequently, they cannot, through their own actions, "beat the system" or corrupt the indicator.

(6) TIMSS provides the international benchmark for the state NAEP results. This occurs only after TIMSS results and NAEP results are expressed in the same metric (percent at and above Proficient)—in other words, after the NAEP-TIMSS linking takes place.

These results give states information on how they perform, not only in comparison to other states, but with other nations throughout the world. This type of information can allow states to not only monitor progress, but also to know how much progress is needed as measured against international benchmarks.

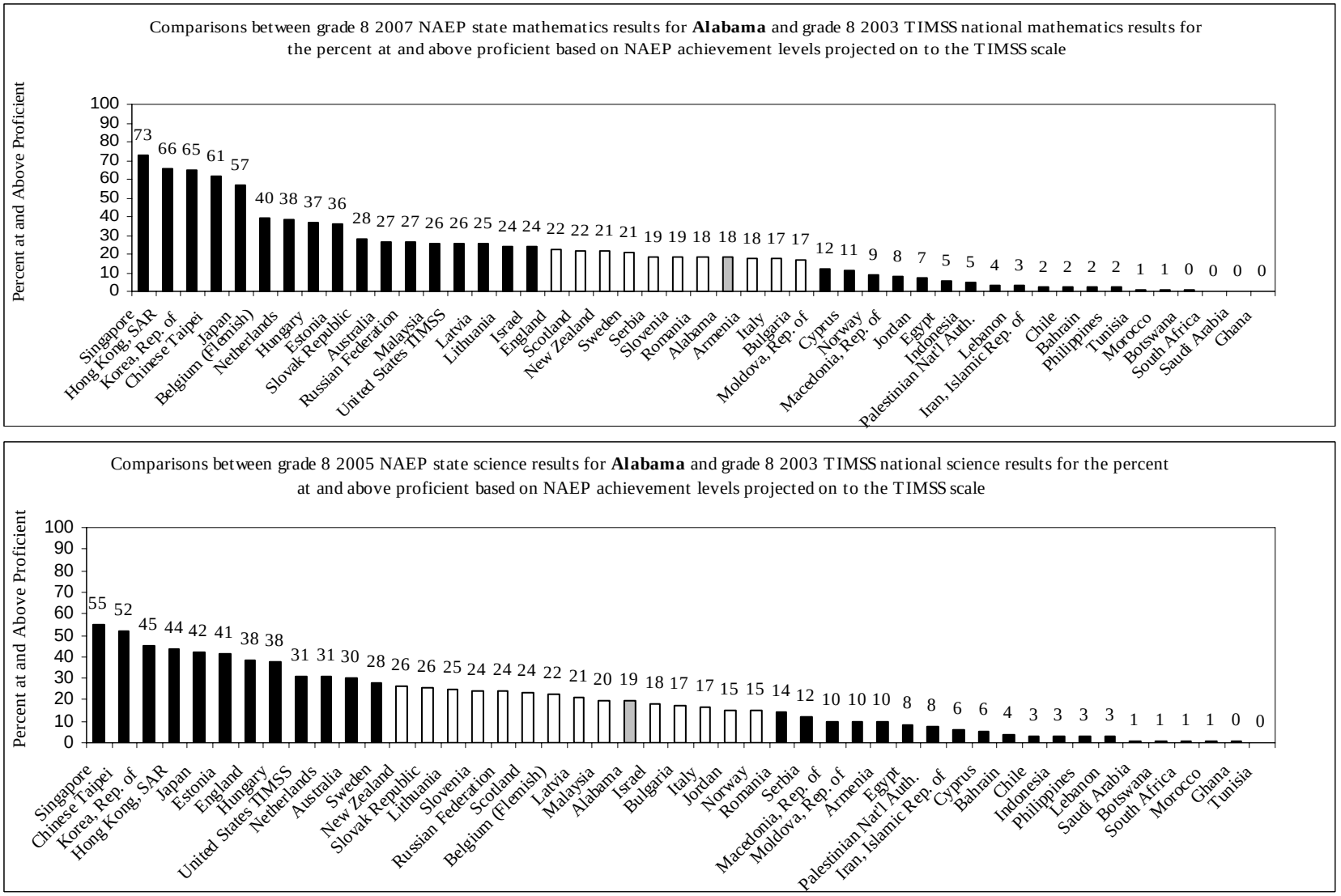
There is an illustrative anecdote that occurred during the 1991 IAEP. One of the monitors that attended both the assessments in South Korea and the United States reported on how the tests were perceived by the students in the two countries. In a U.S. school, students were taken to the cafeteria and a subset was randomly selected for the assessment. The students selected were laughed at because of their bad luck at having to take the test. In a South Korean school, the same procedure was repeated, but the students were cheered for their good luck at the chance to represent their country.

More than a century ago, Louis Pasteur revealed the secret to scientific invention and innovation when he said, "Chance favors the prepared mind." How well have we prepared the minds of our students to improve their chances? The results in this report represent both good news and bad news. The *good news* is that most states are doing as well or better than most foreign countries. *If you think of states and nations as in a race to prepare the future generation of workers, scholars and citizens to be competent and competitive in a technologically complex world, then the states are in the middle of the pack.* The *bad news* is that even our best-performing states are significantly below the highest performing countries.

This report shows that the American public has very low levels of mathematical and scientific literacy. Instead of relying on science we rely on pseudoscience. Our public school students are not keeping up with their Asian counterparts who will be their economic competitors in the future. Our colleges are not graduating enough students in the scientific and engineering fields today that would provide the advances in technology needed for tomorrow. The take away message from this report is that the United States is loosing the race to prepare the minds of the future generation.

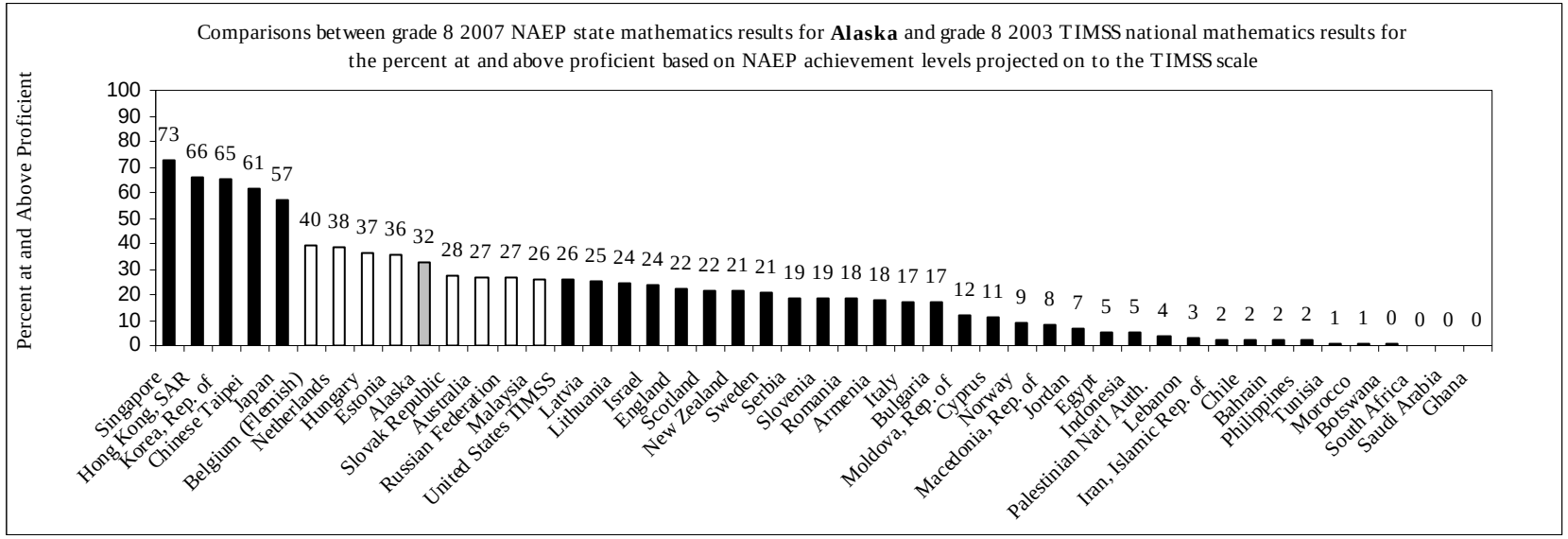
Figures showing each state in NAEP compared to each nation in TIMSS

Figure 1: Alabama



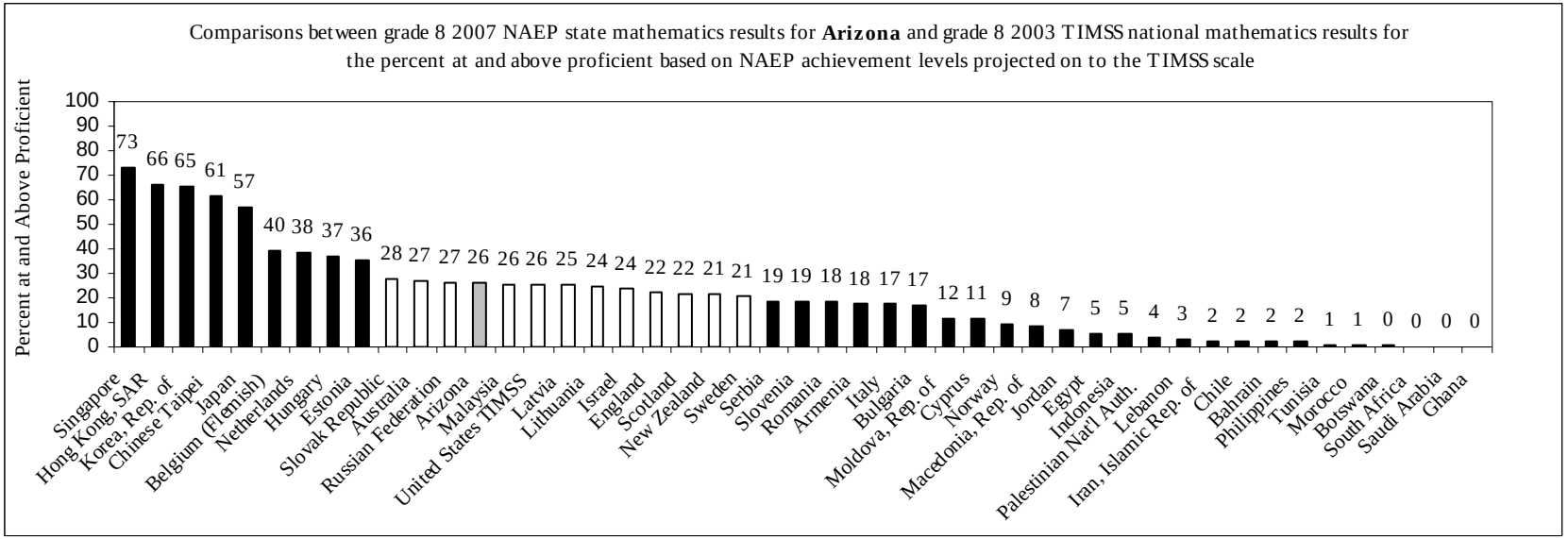
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

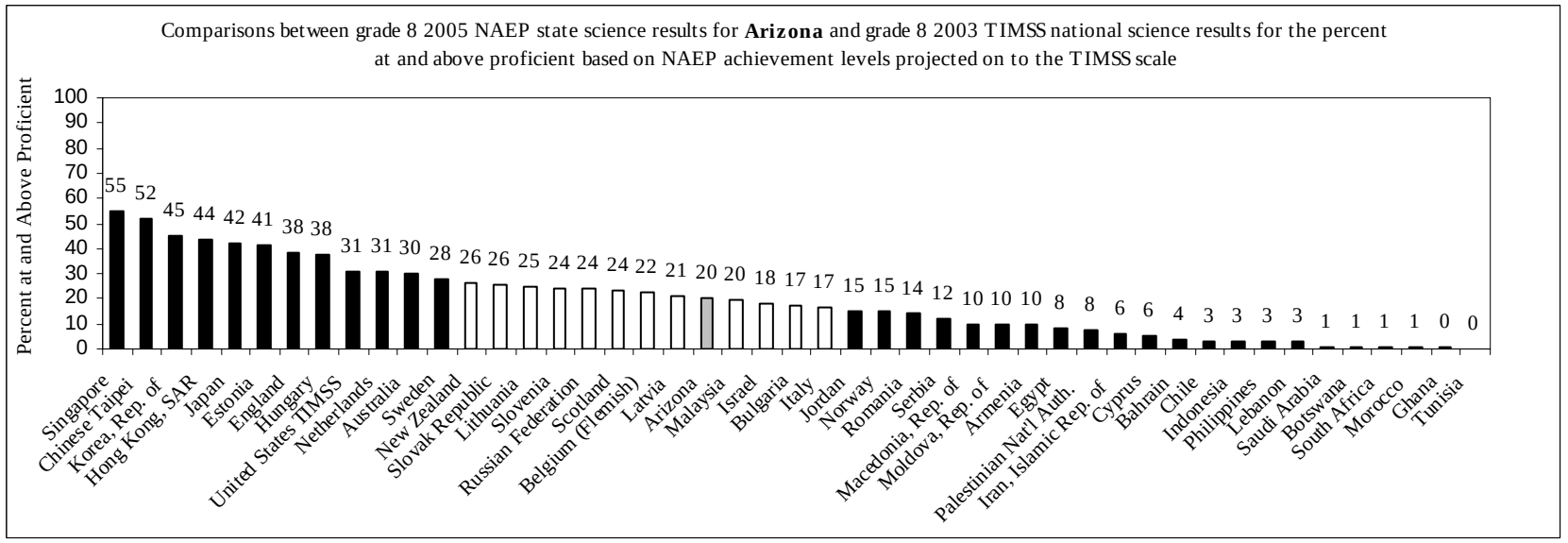
Figure 2: Alaska



Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.
Alaska did not participate in the grade 8 2005 state NAEP in science.

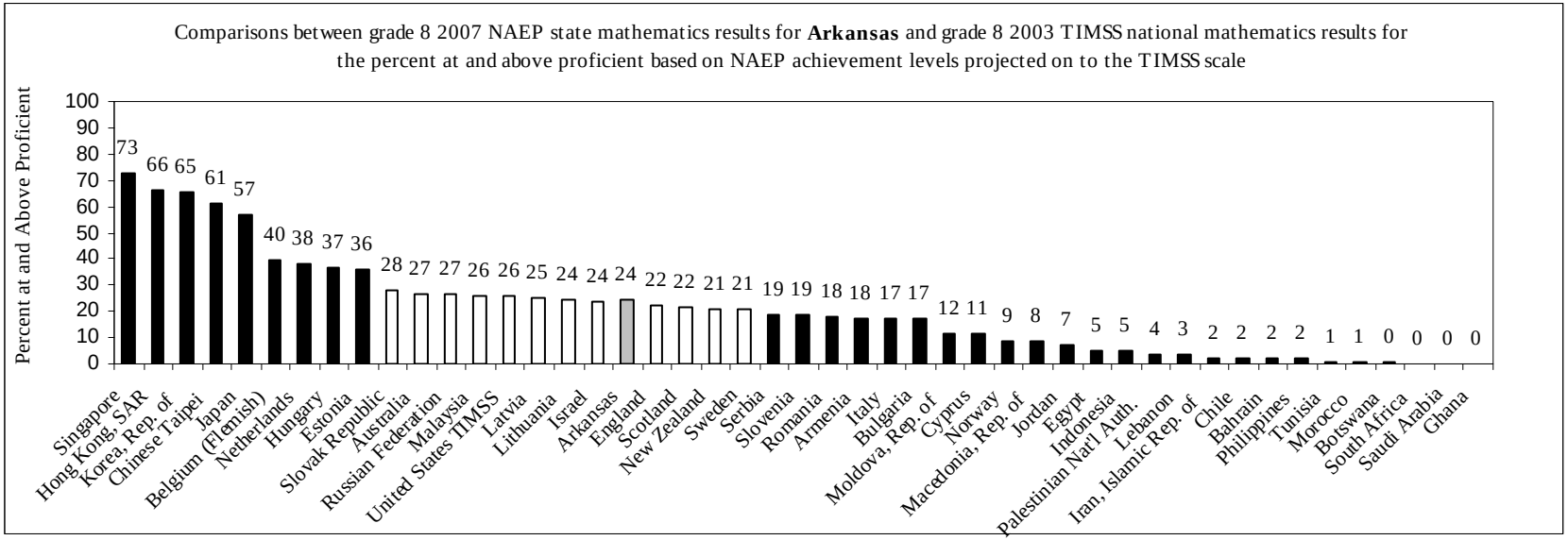
Figure 3: Arizona

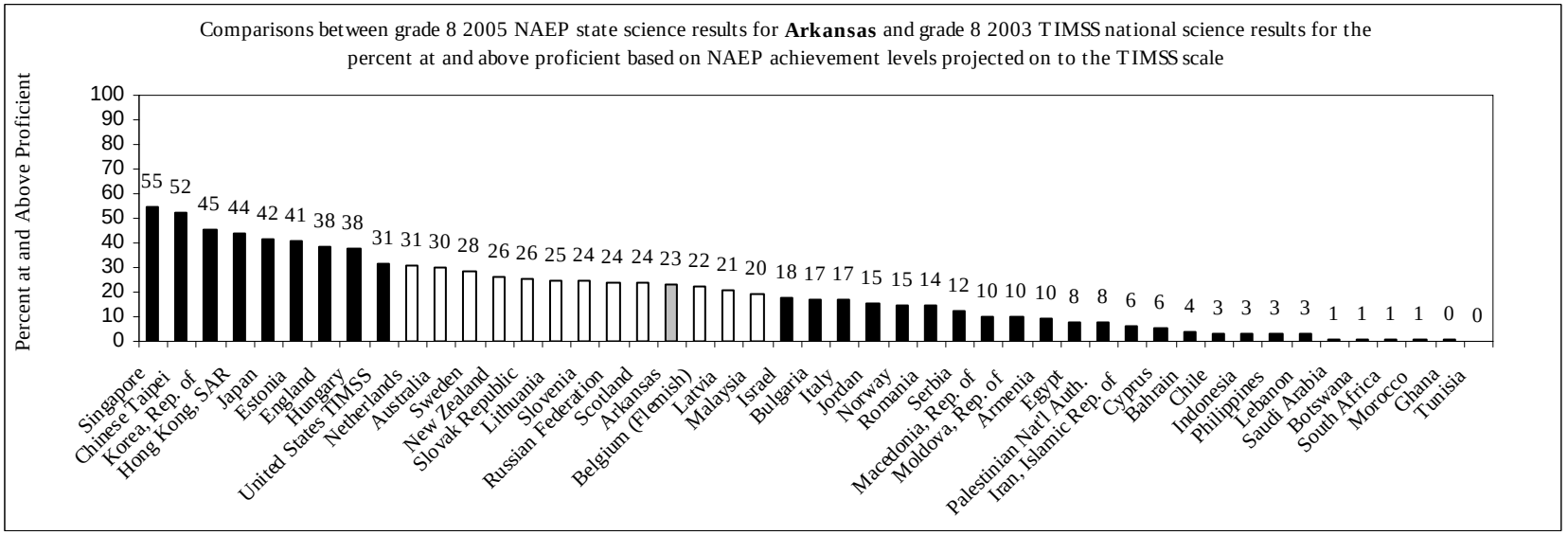




Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

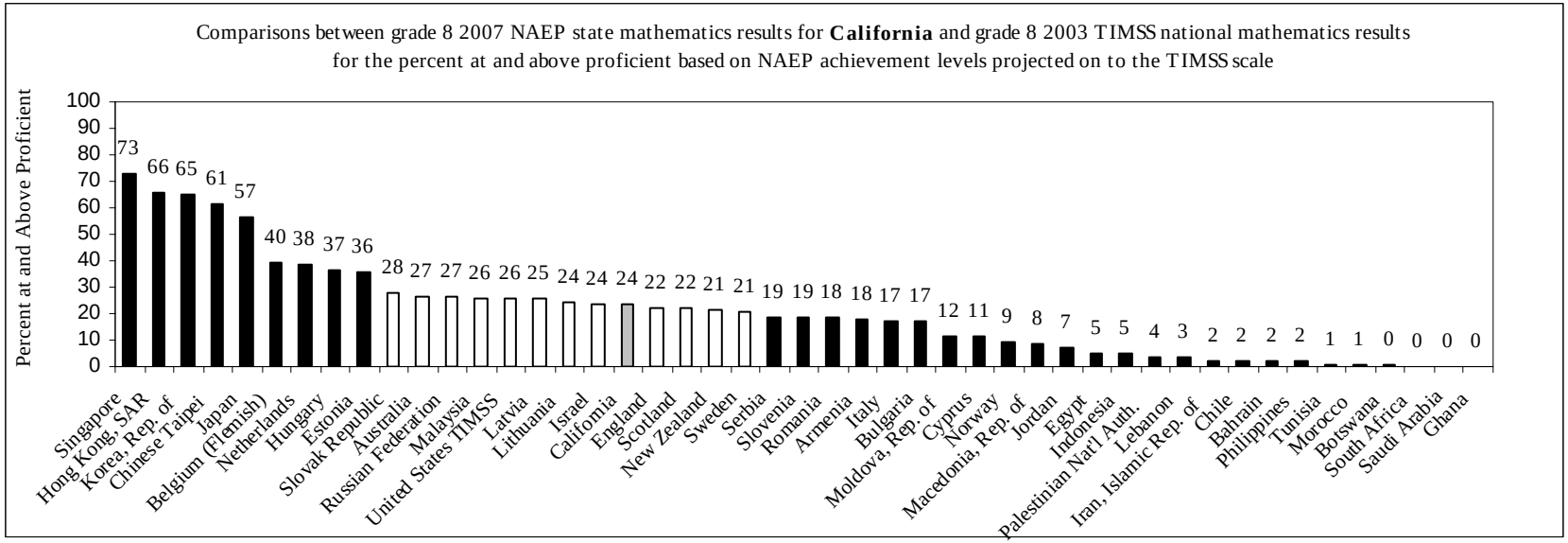
Figure 4: Arkansas

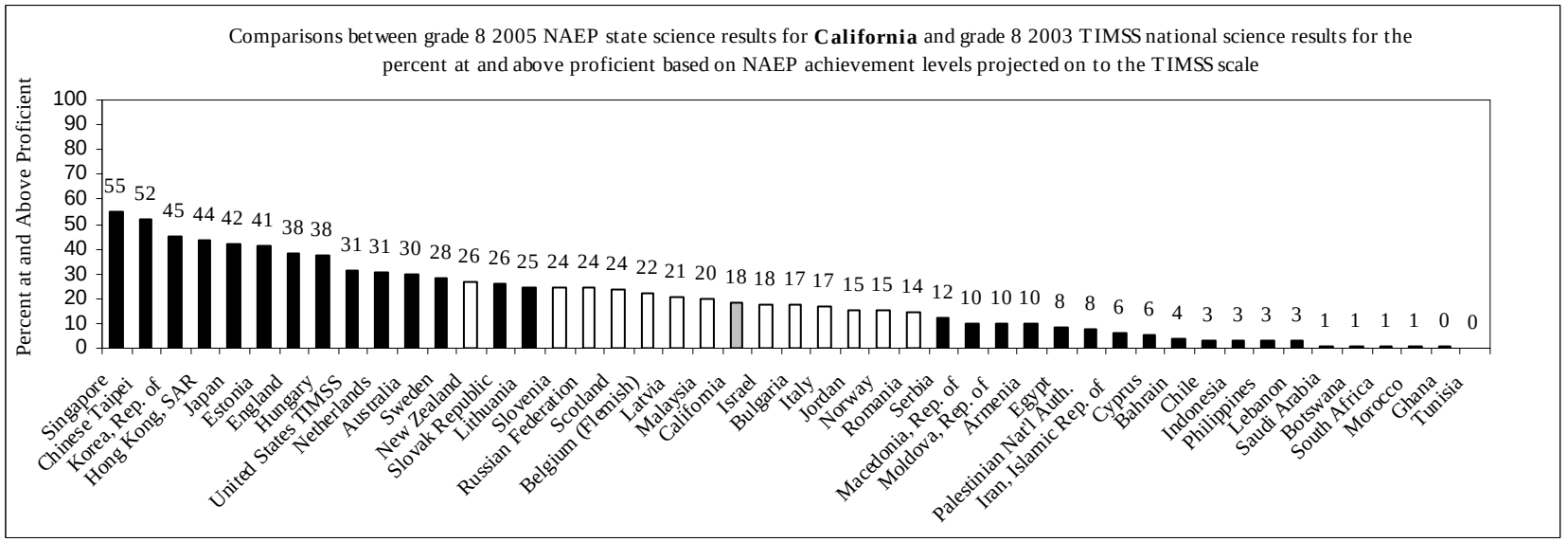




Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

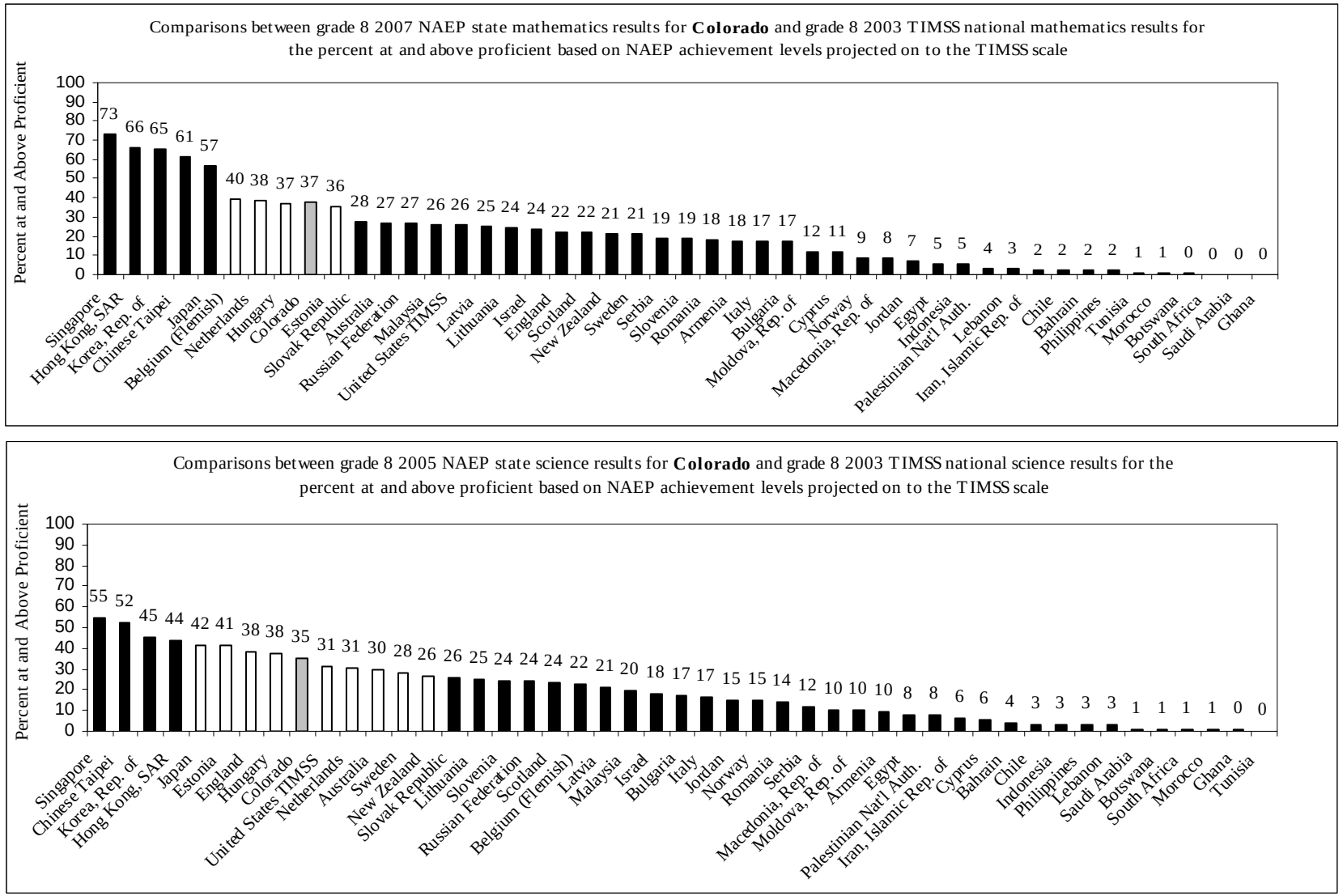
Figure 5: California





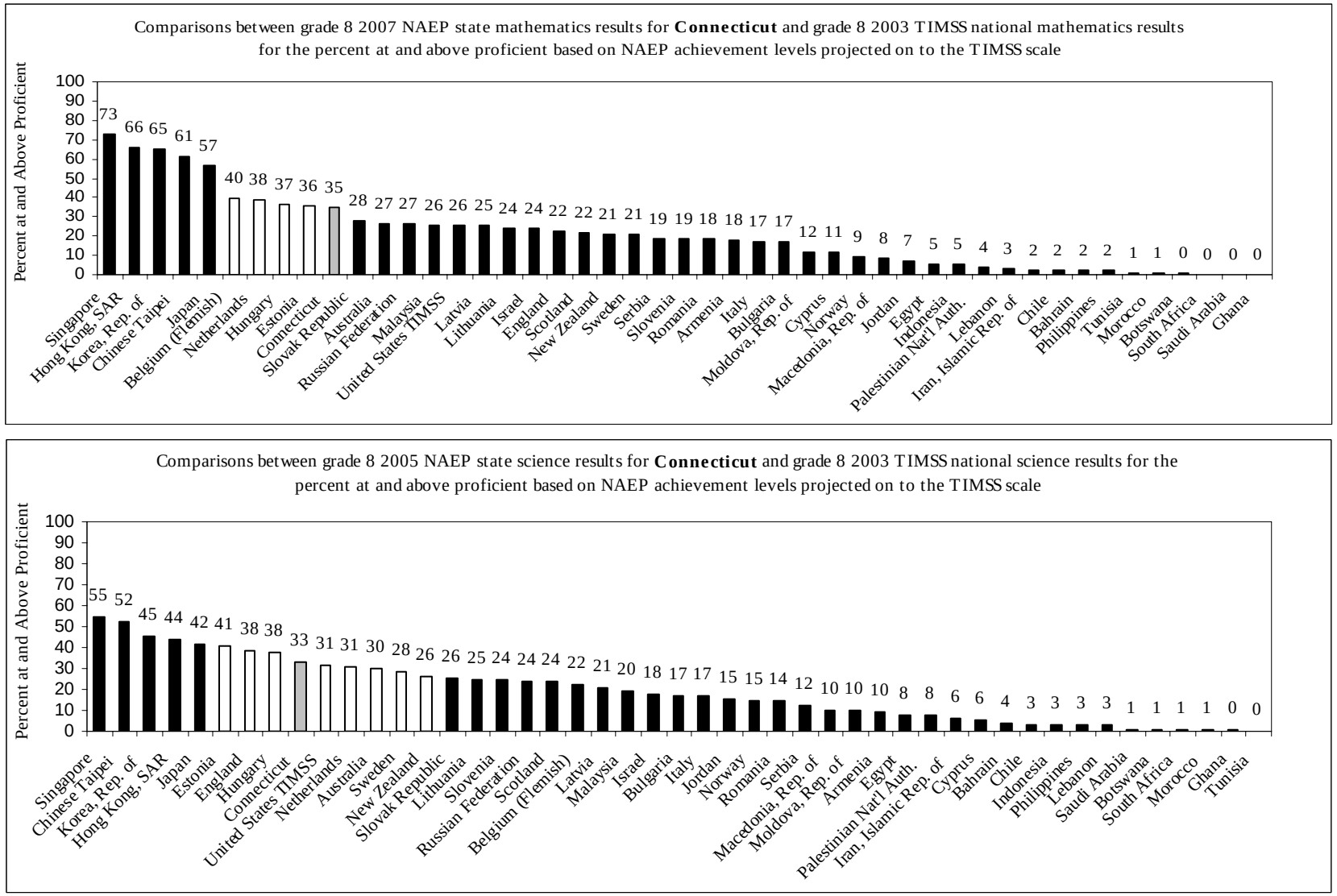
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

Figure 6: Colorado



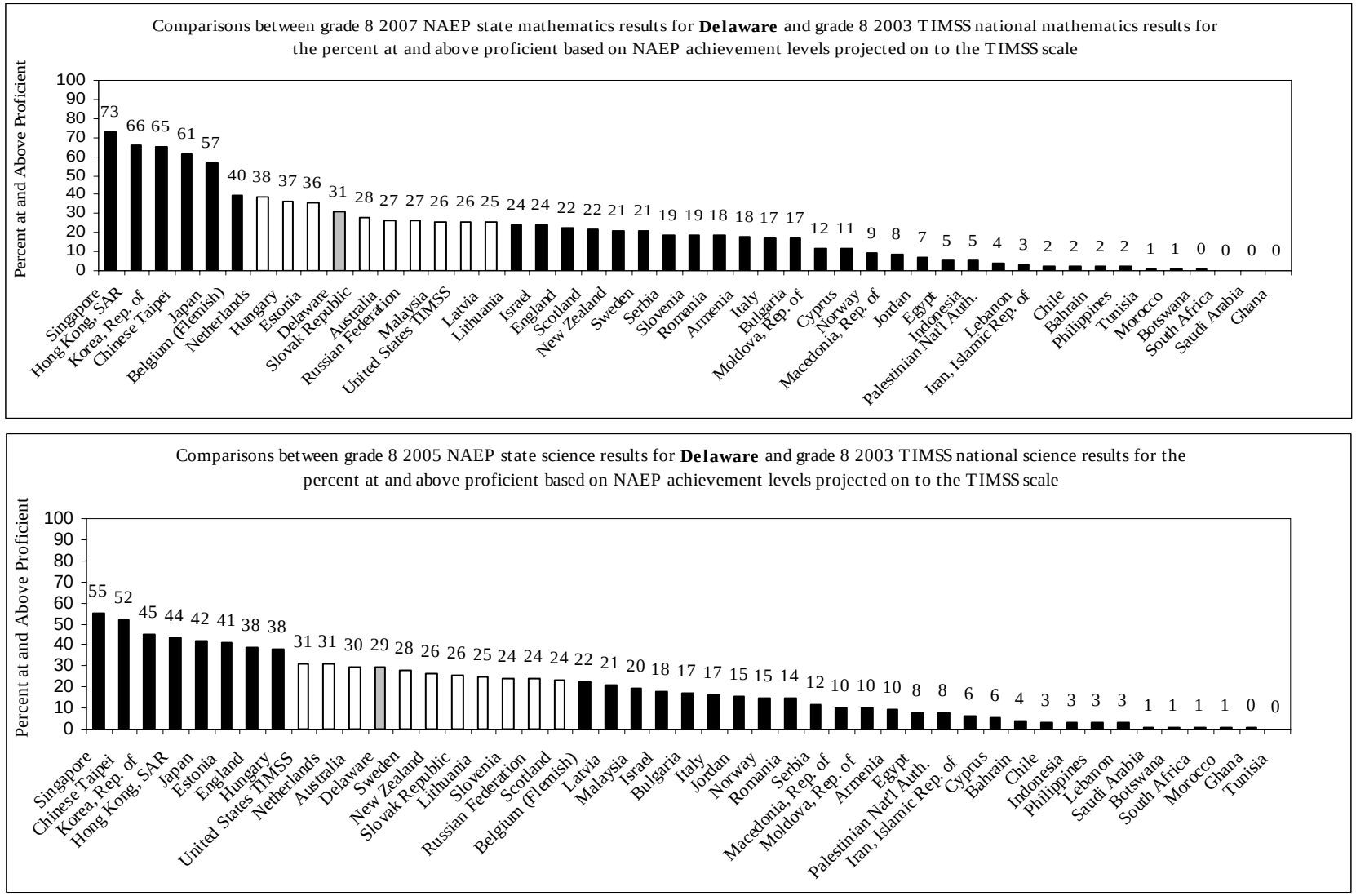
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

Figure 7: Connecticut



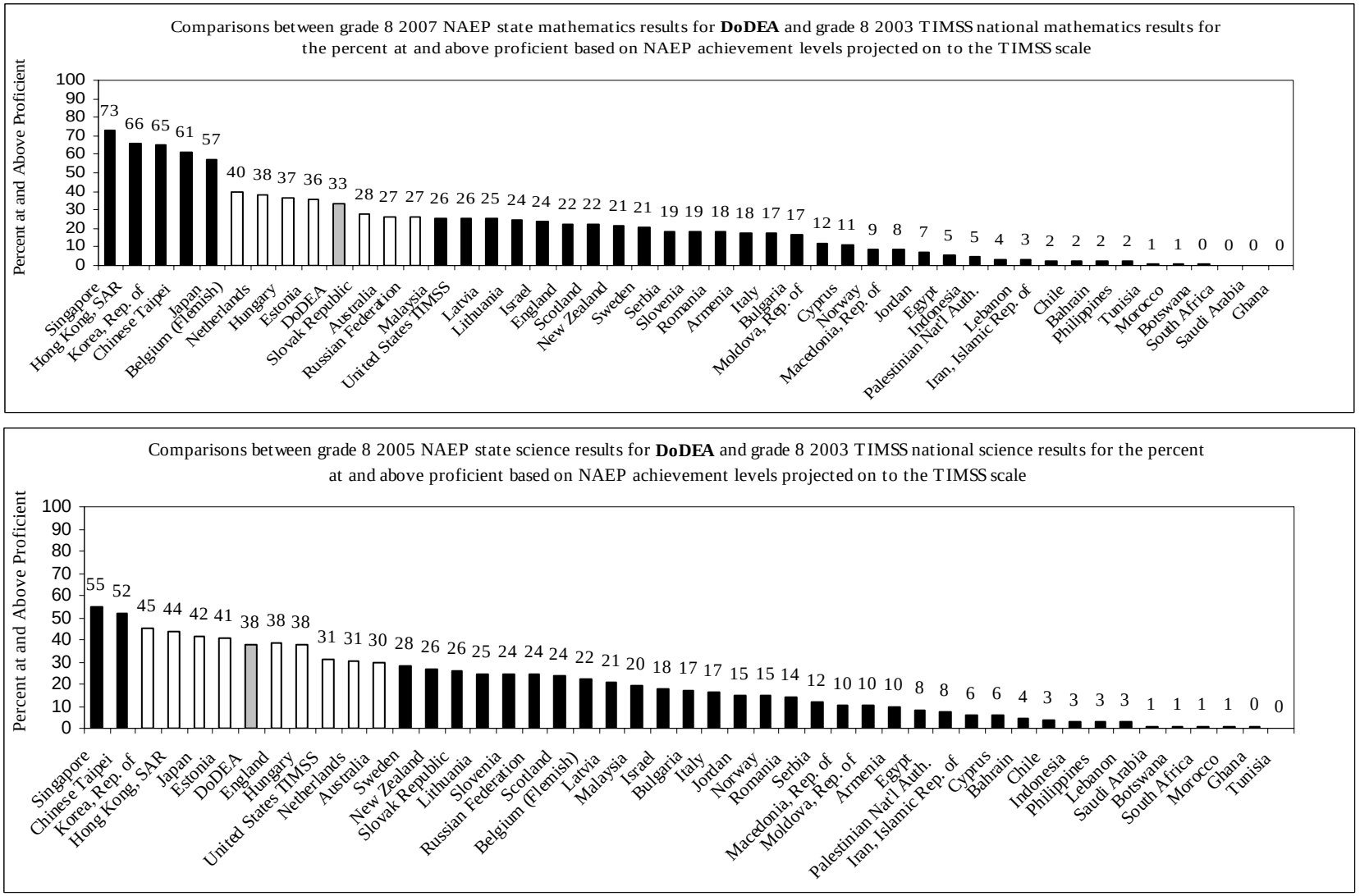
***Source**: Phillips, Gary W., Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations, AIR: Wash., DC, 2007.*

Figure 8: Delaware



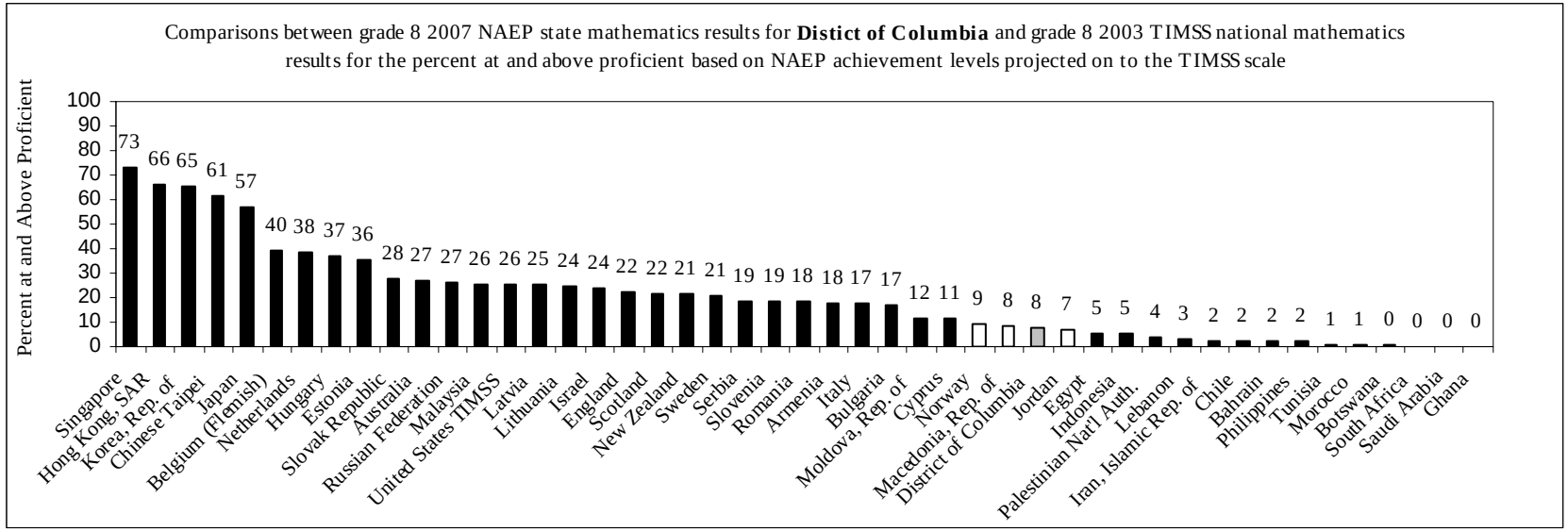
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

Figure 9: Department of Defense Education Activity (DoDEA)



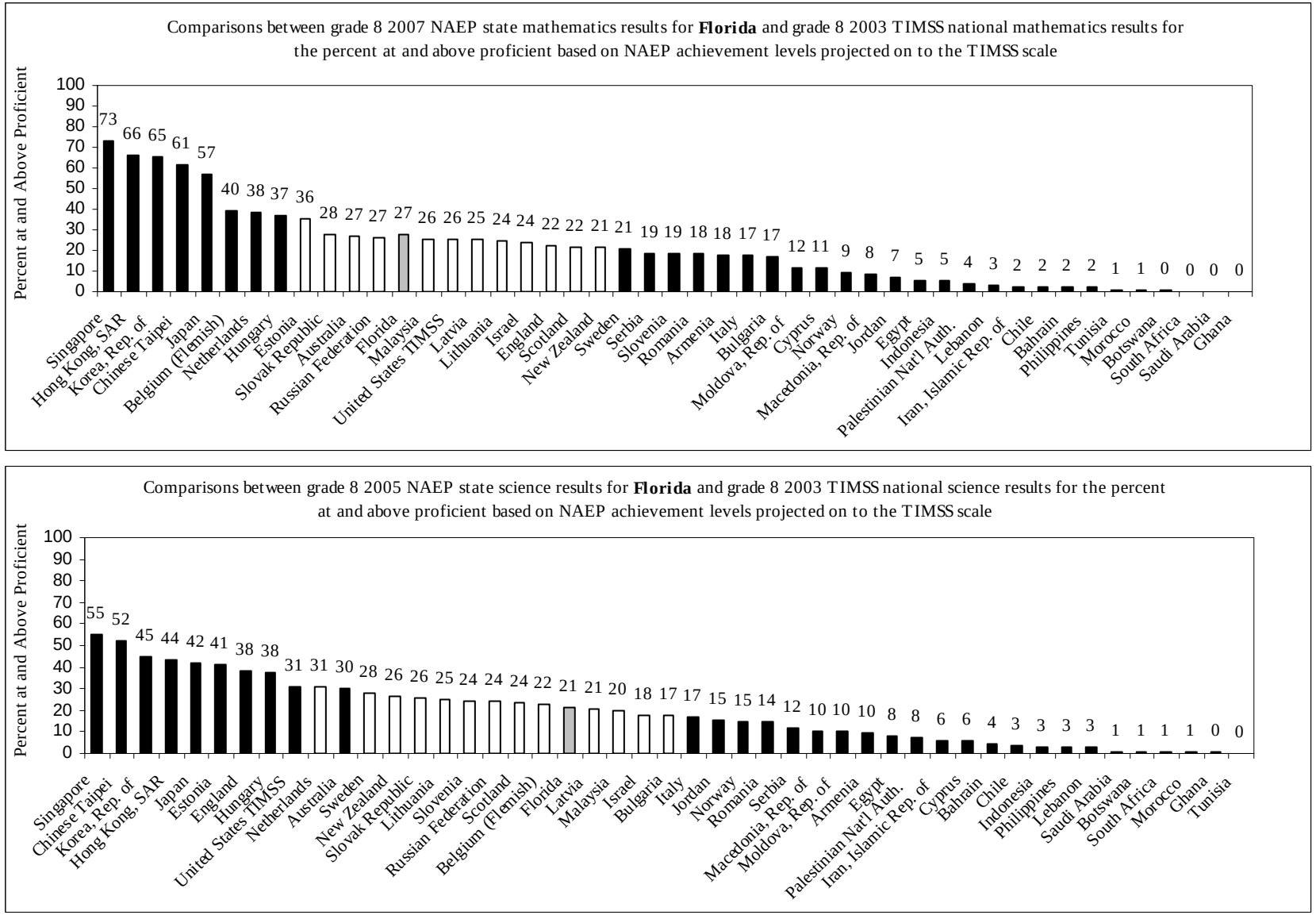
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

Figure 10: District of Columbia



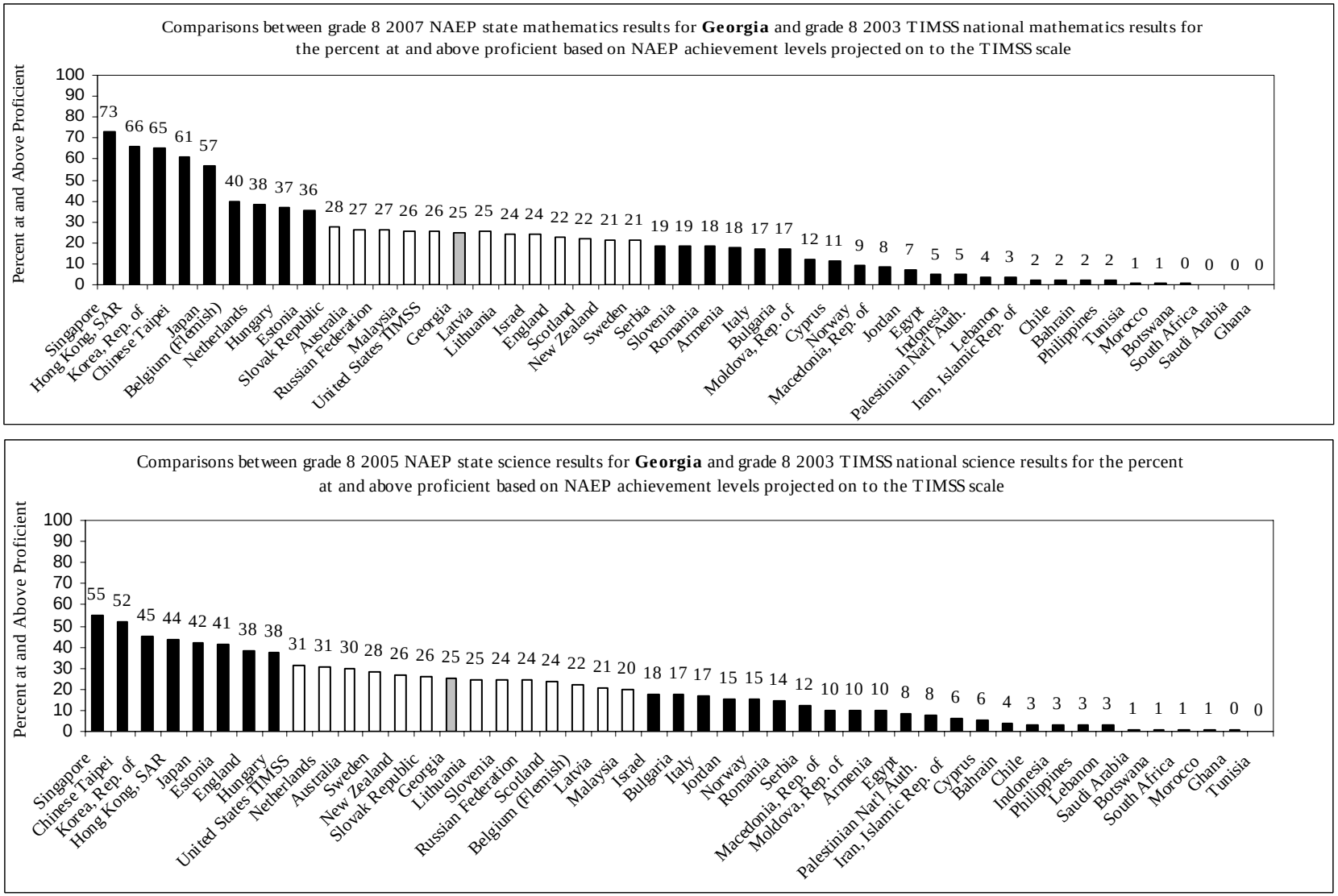
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.
District of Columbia did not participate in the grade 8 2005 state NAEP in science.

Figure 11: Florida



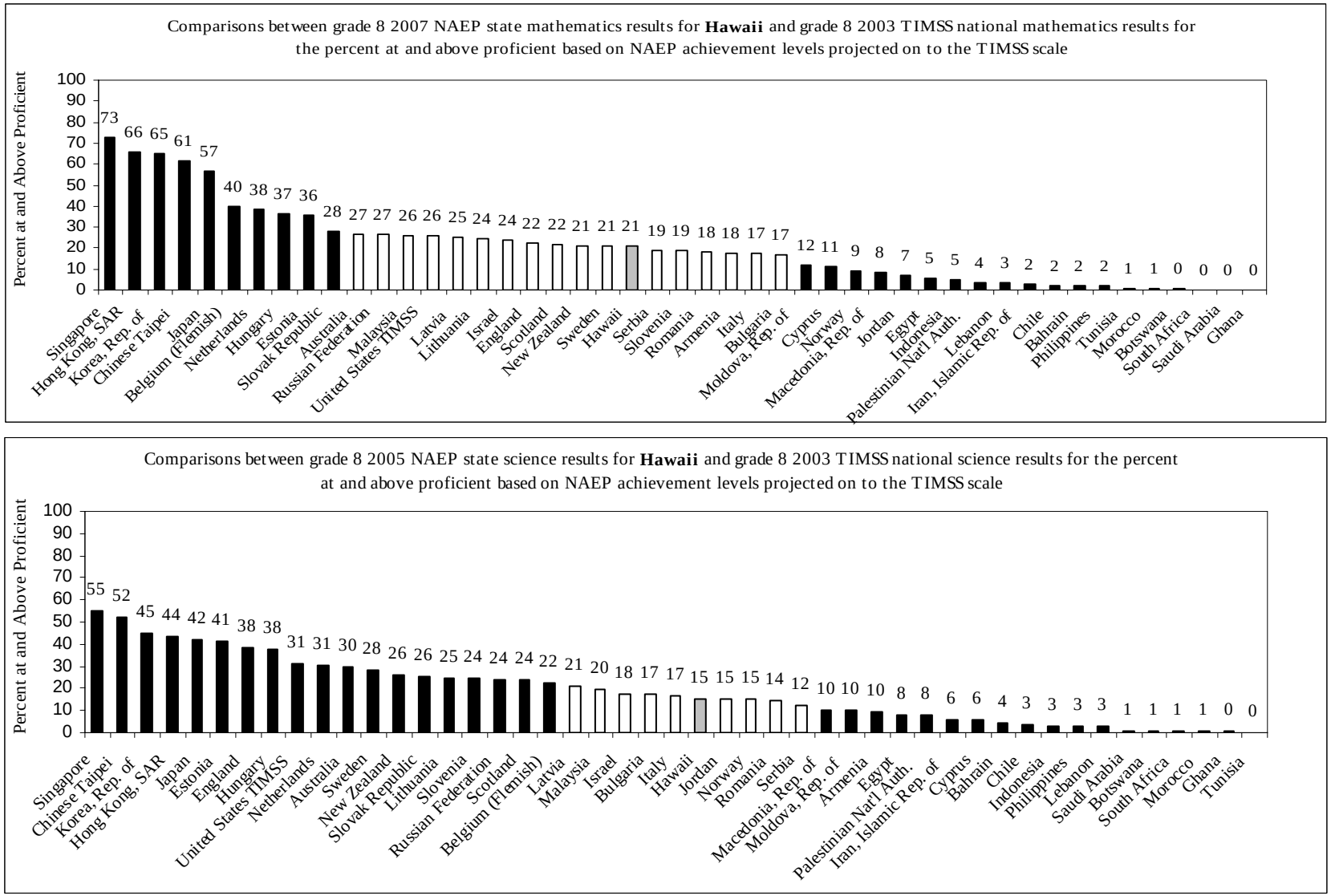
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

Figure 12: Georgia



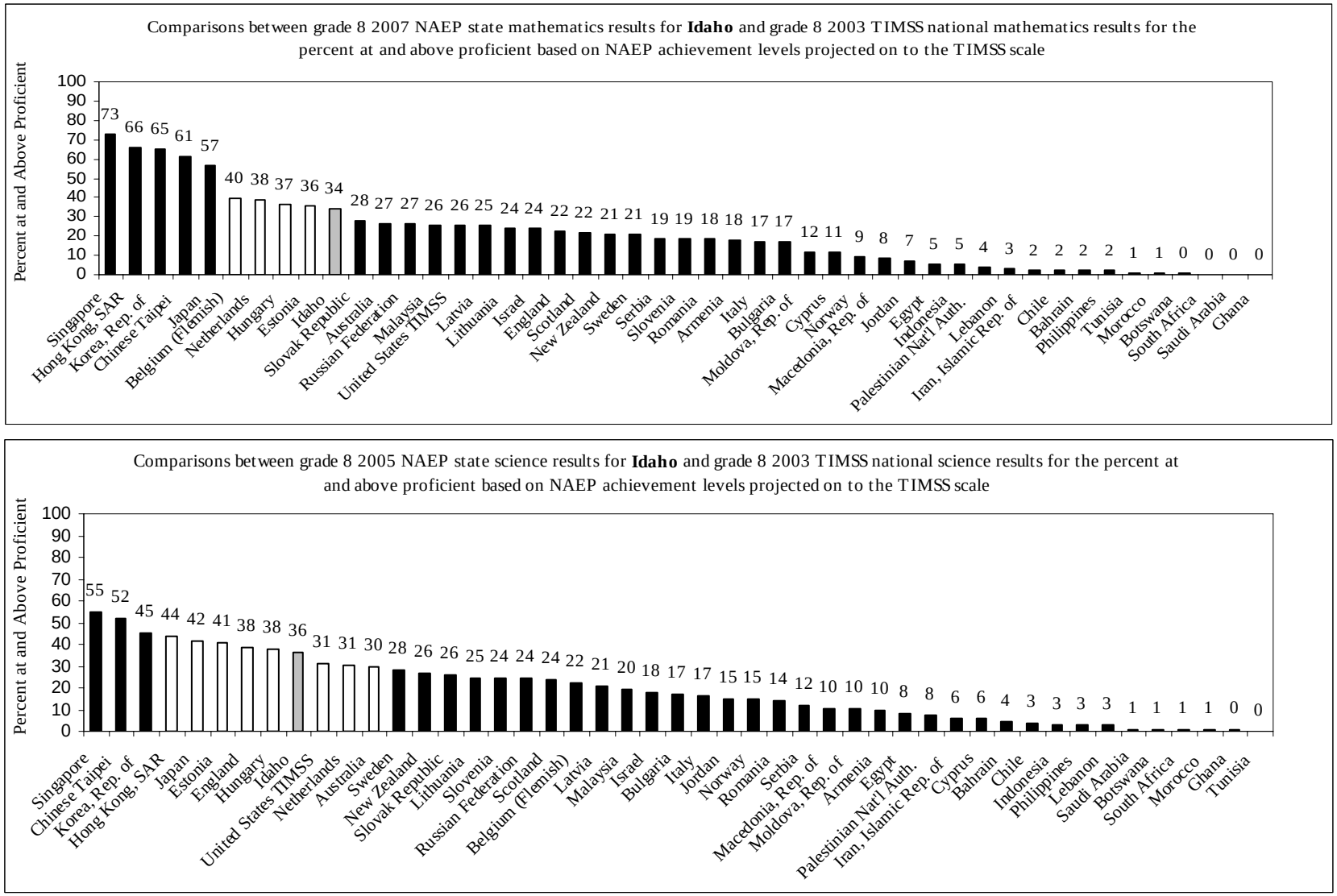
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations,* AIR: Wash., DC, 2007**.**

Figure 13: Hawaii



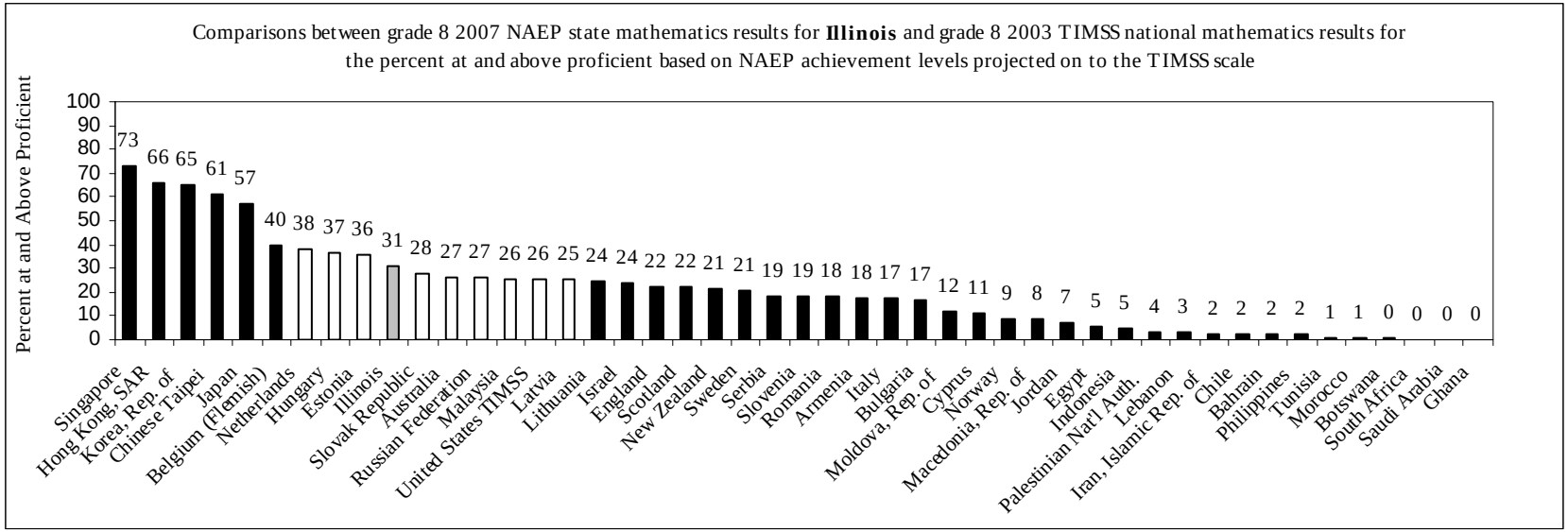
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

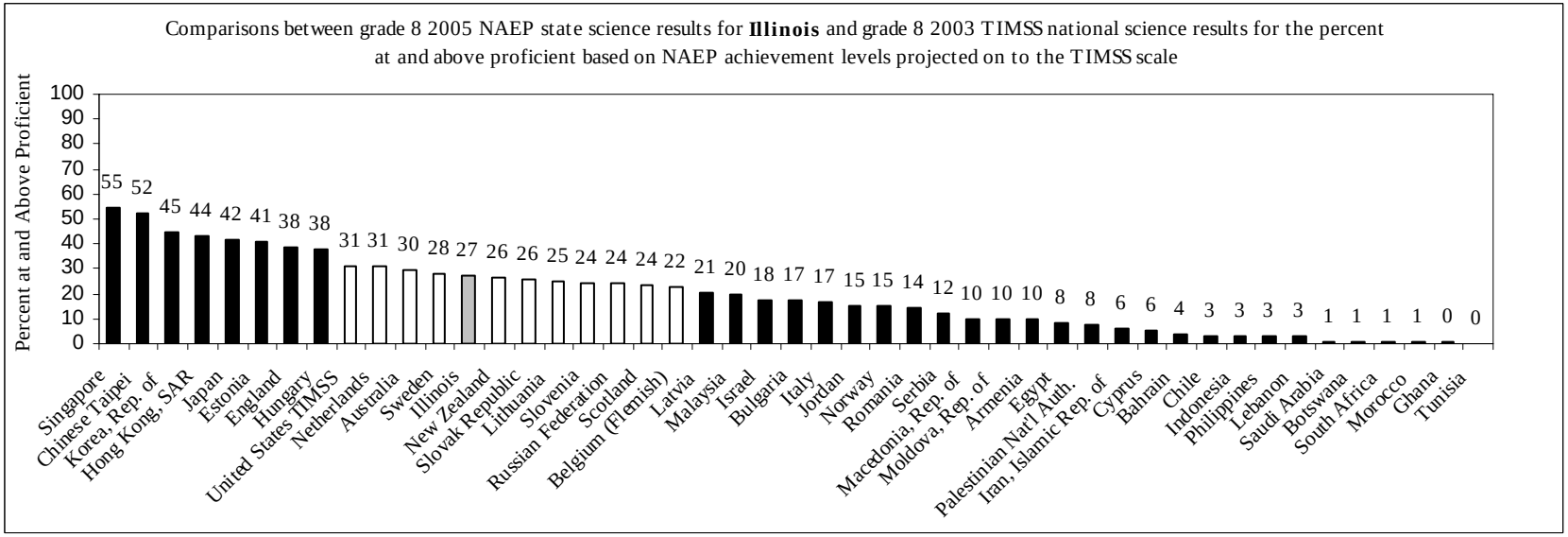
Figure 14: Idaho



Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

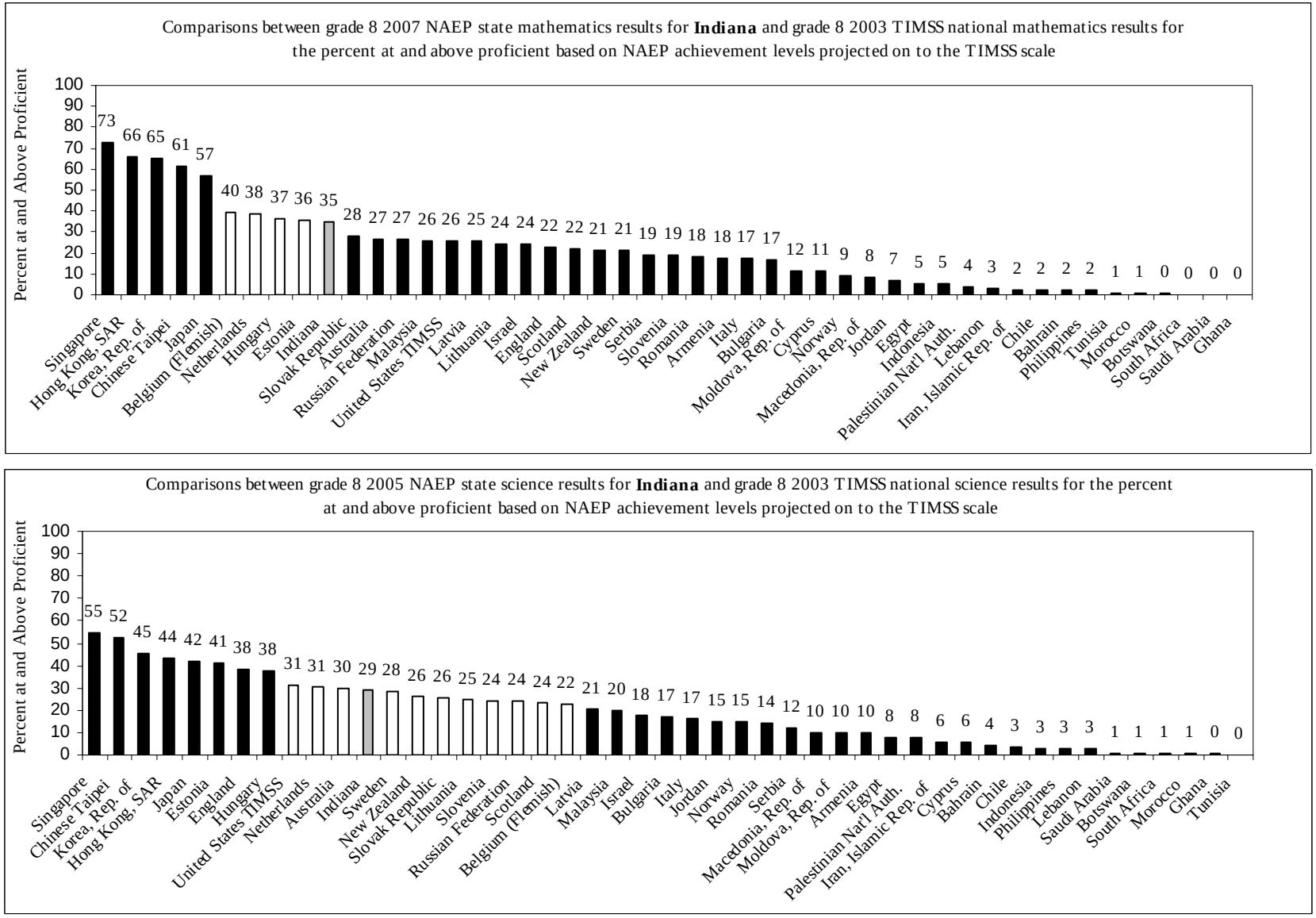
Figure 15: Illinois





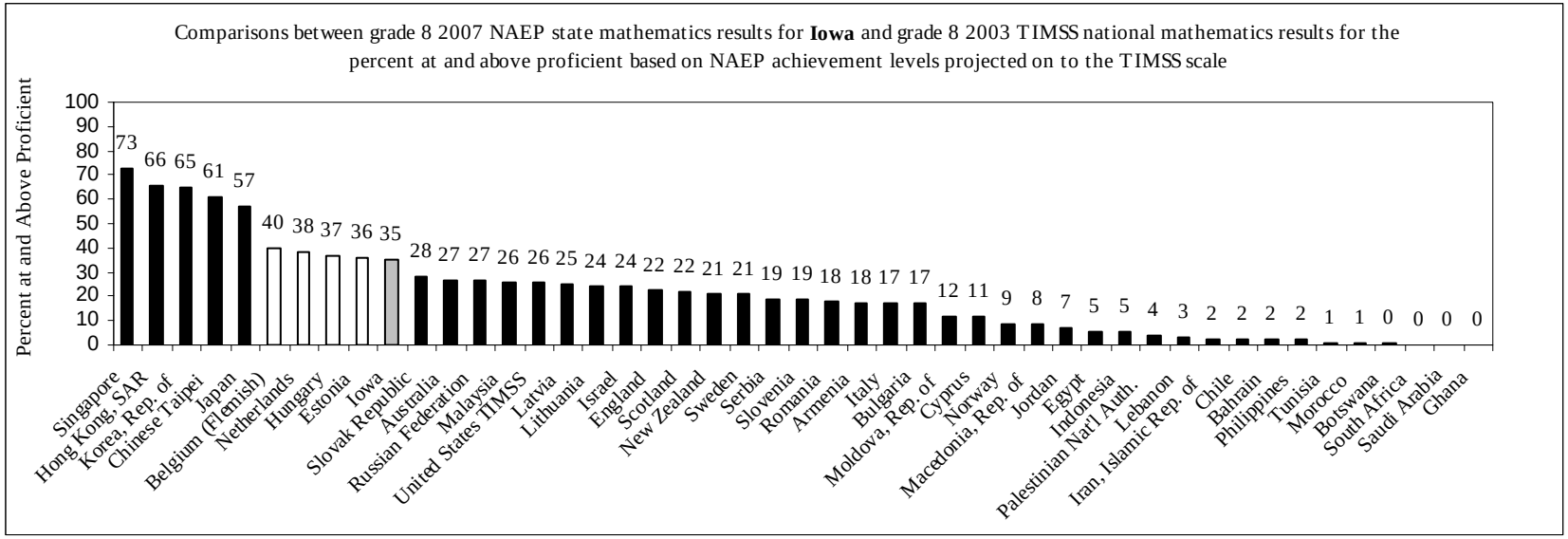
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

Figure 16: Indiana



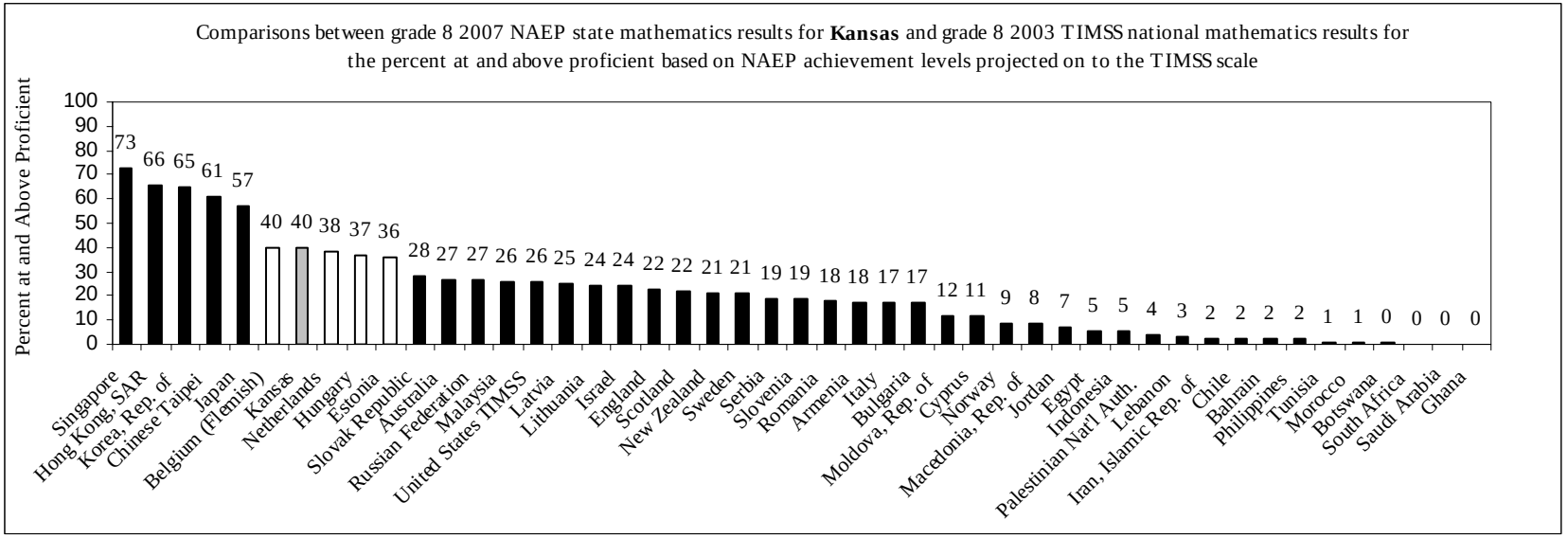
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

Figure 17: Iowa



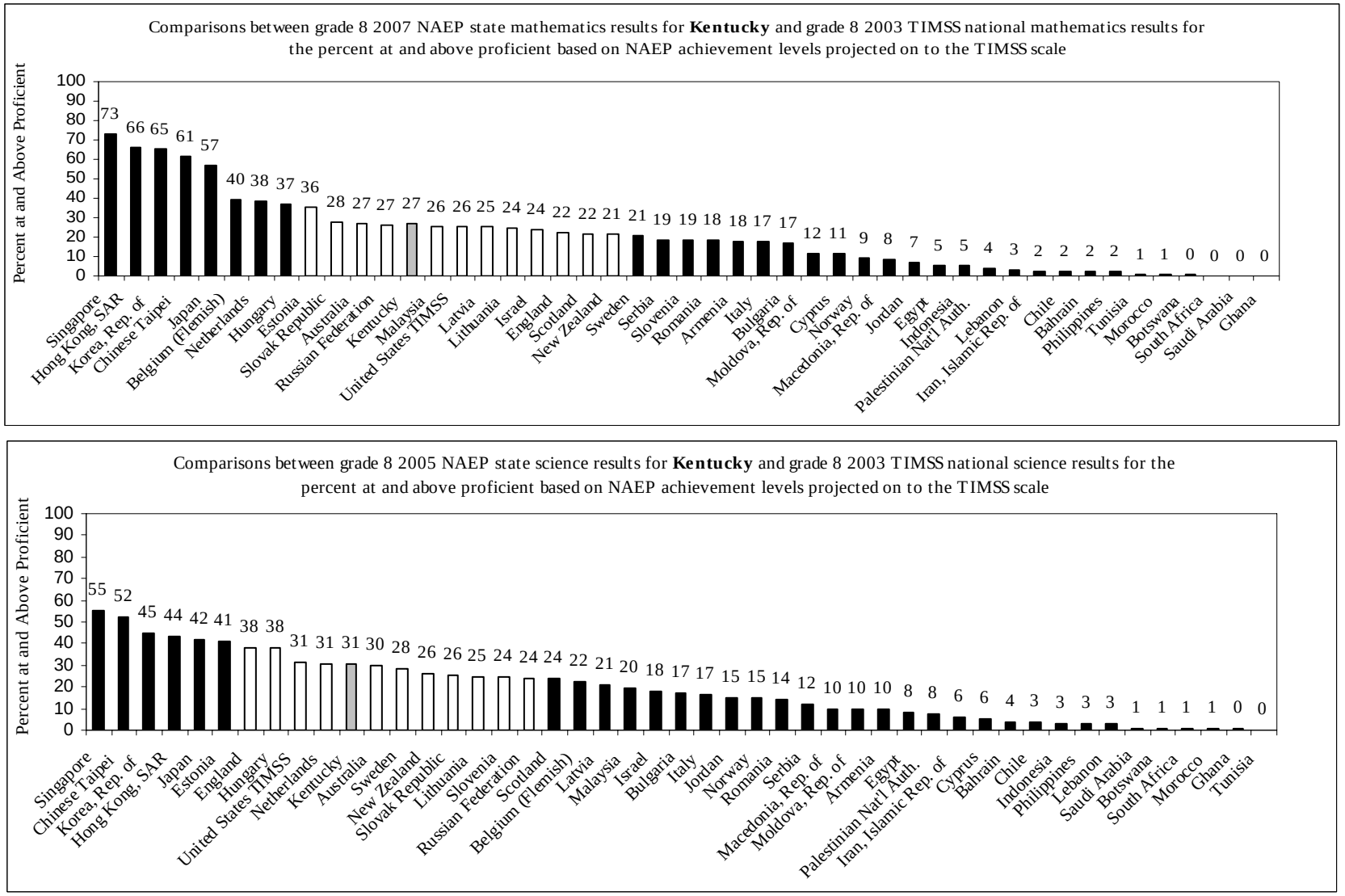
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.
Iowa did not participate in the grade 8 2005 state NAEP in science.

Figure 18: Kansas



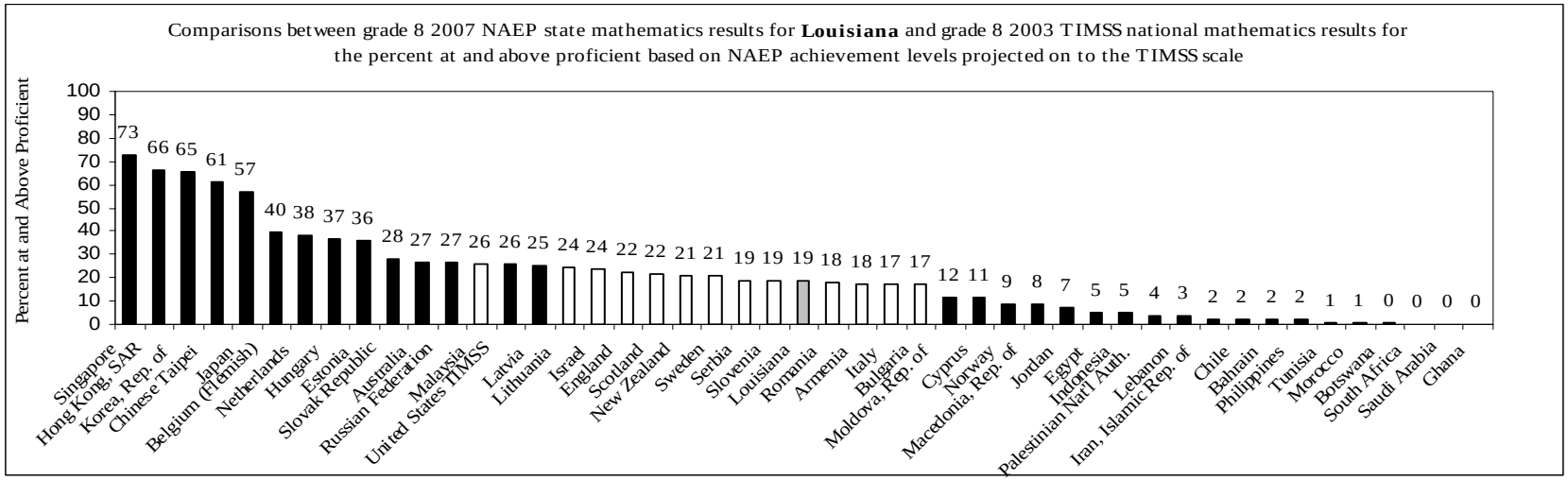
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 20072007.
Kansas did not participate in the grade 8 2005 state NAEP in science.

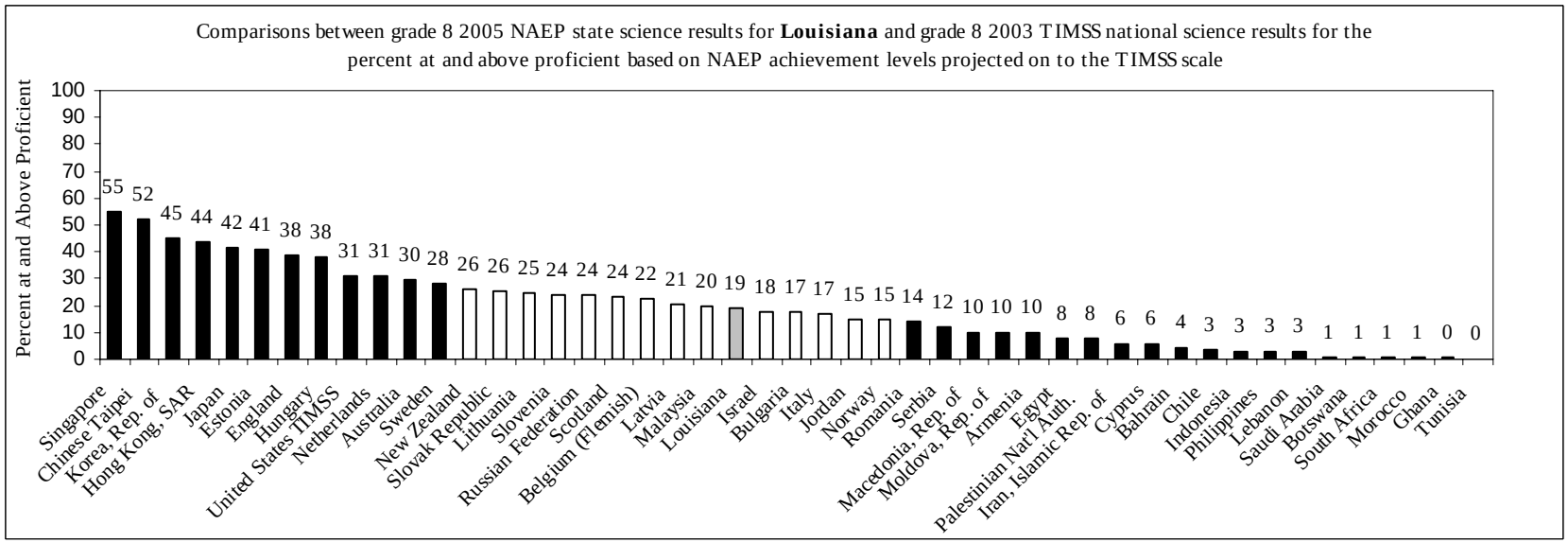
Figure 19: Kentucky



Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

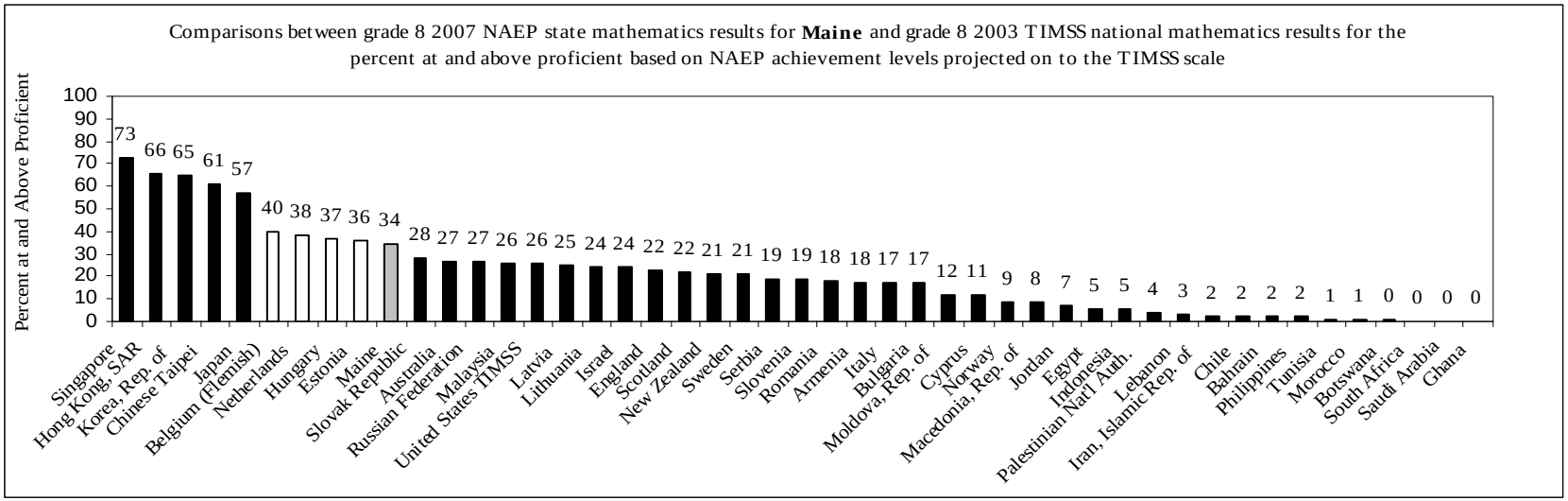
Figure 20: Louisiana

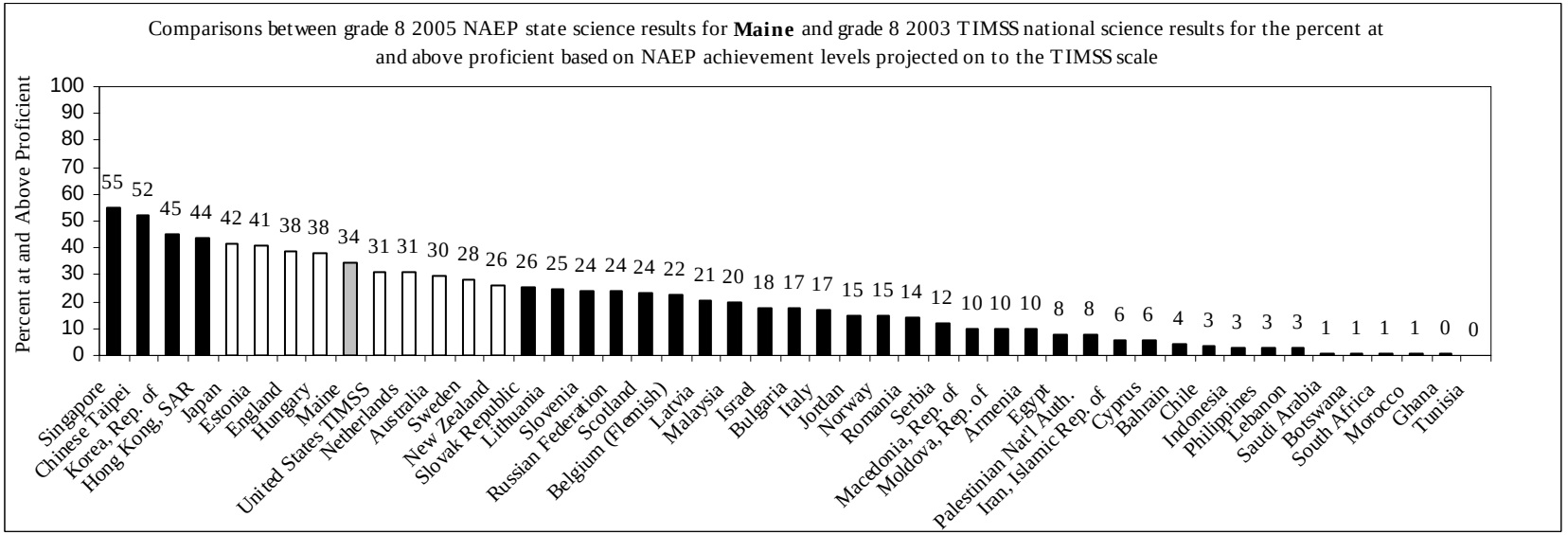




Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

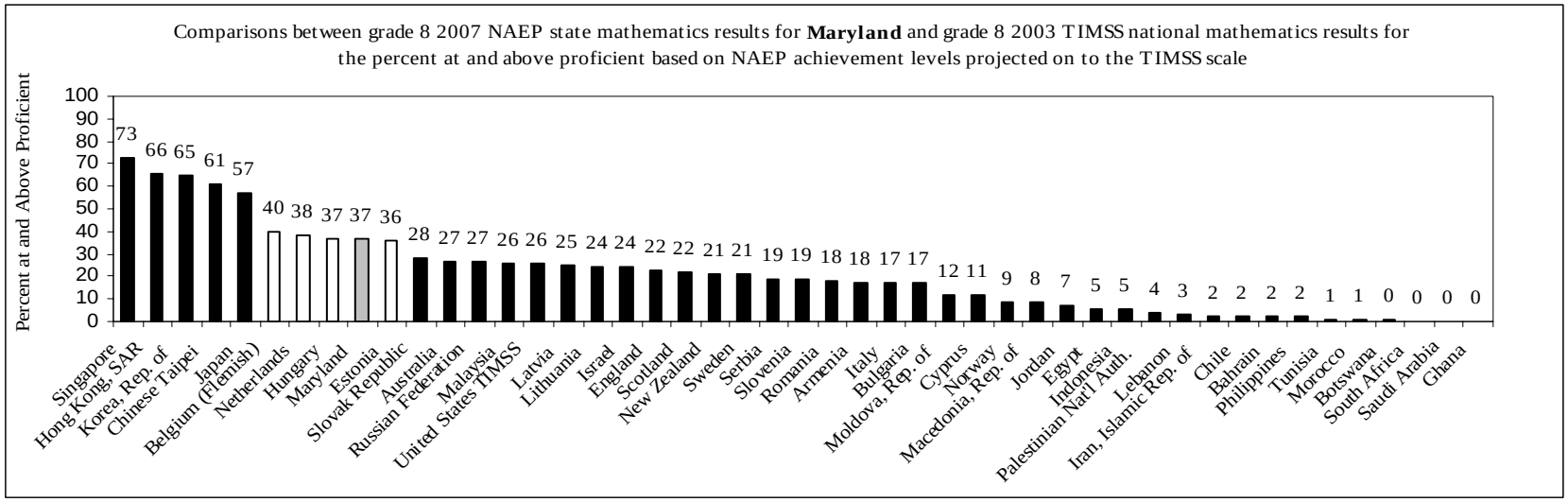
Figure 21: Maine

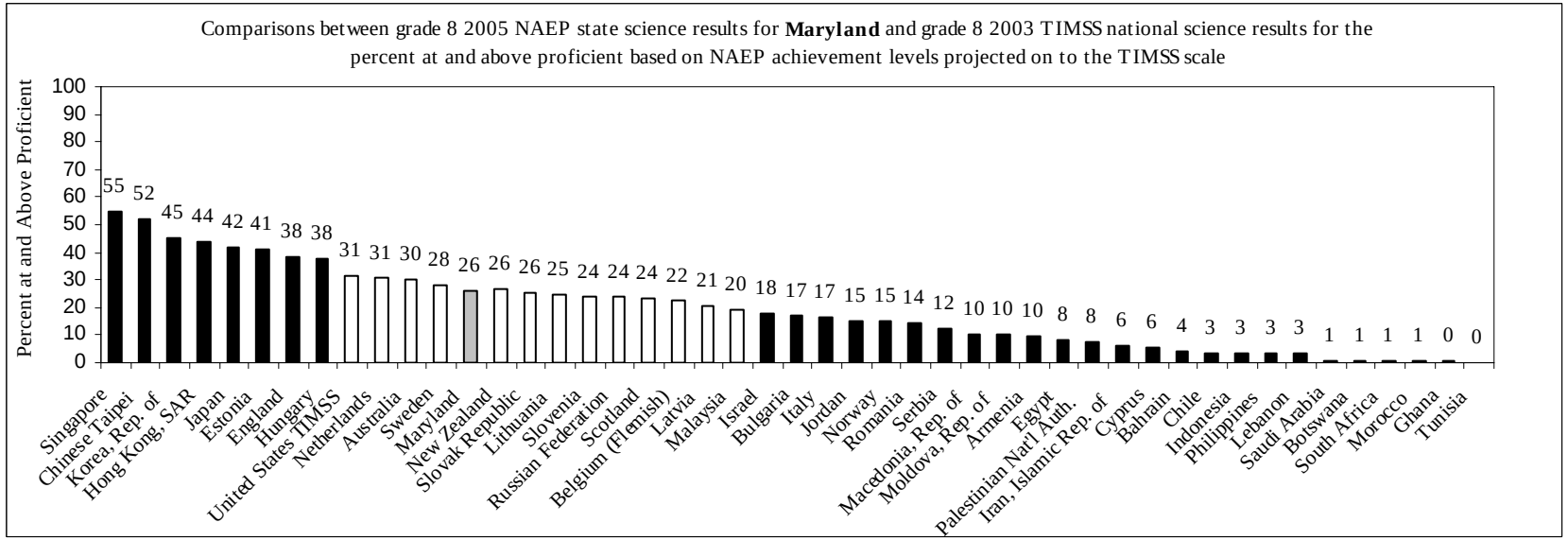




Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

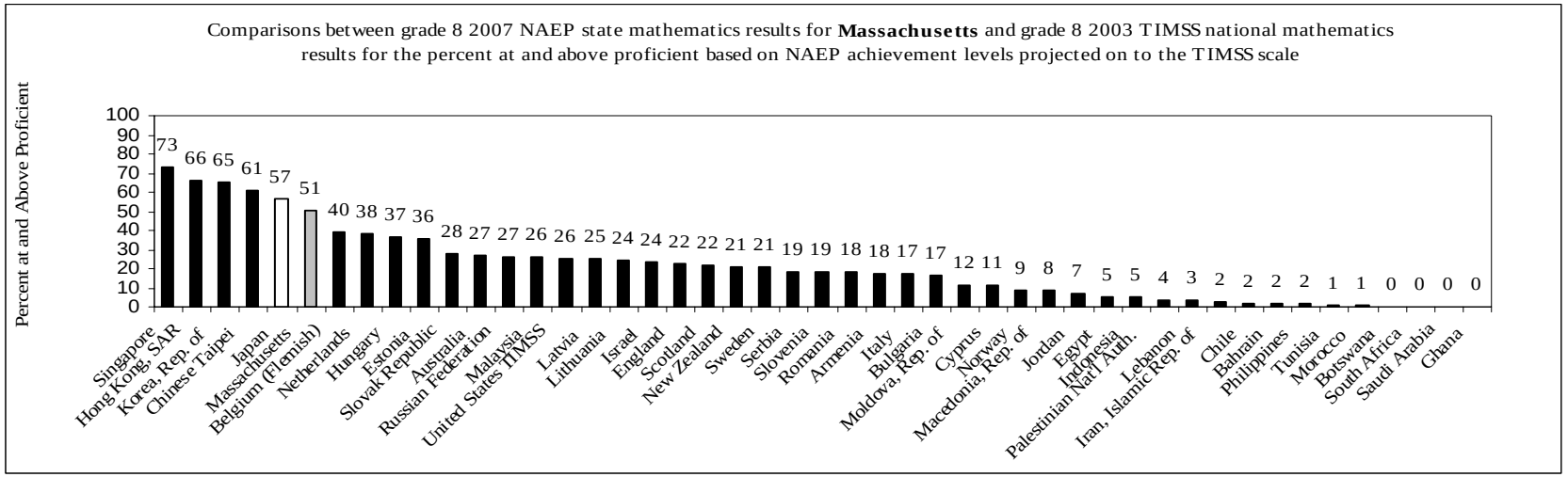
Figure 22: Maryland

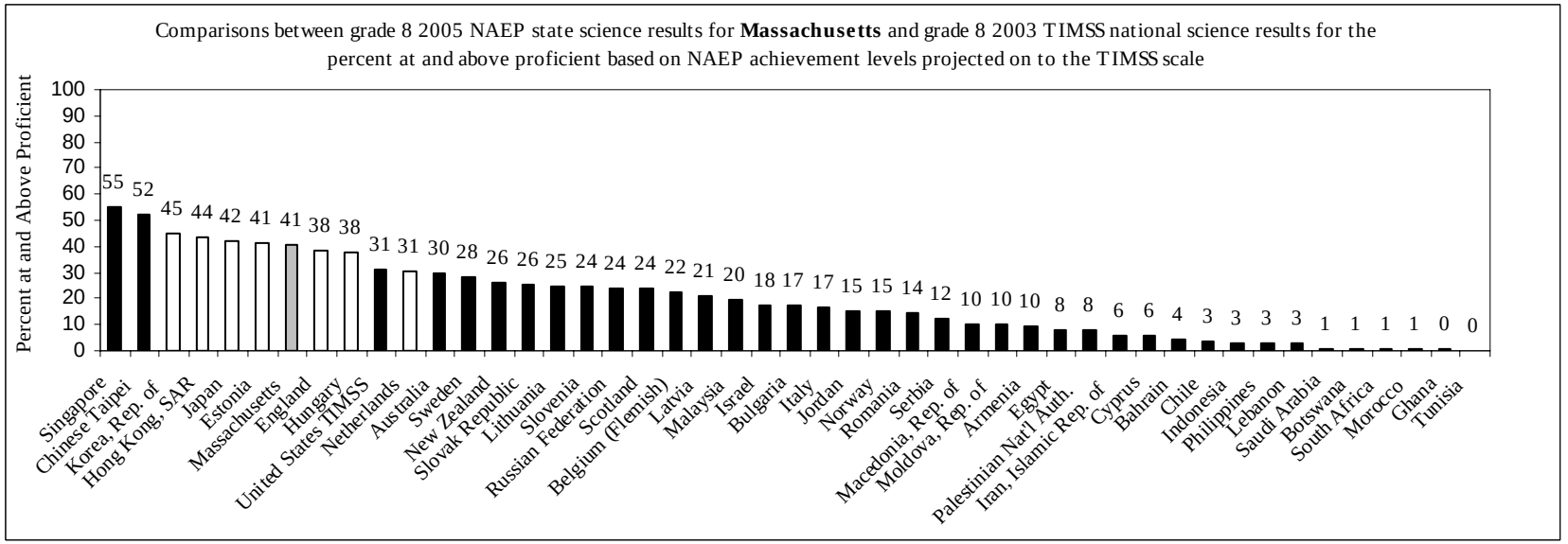




Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

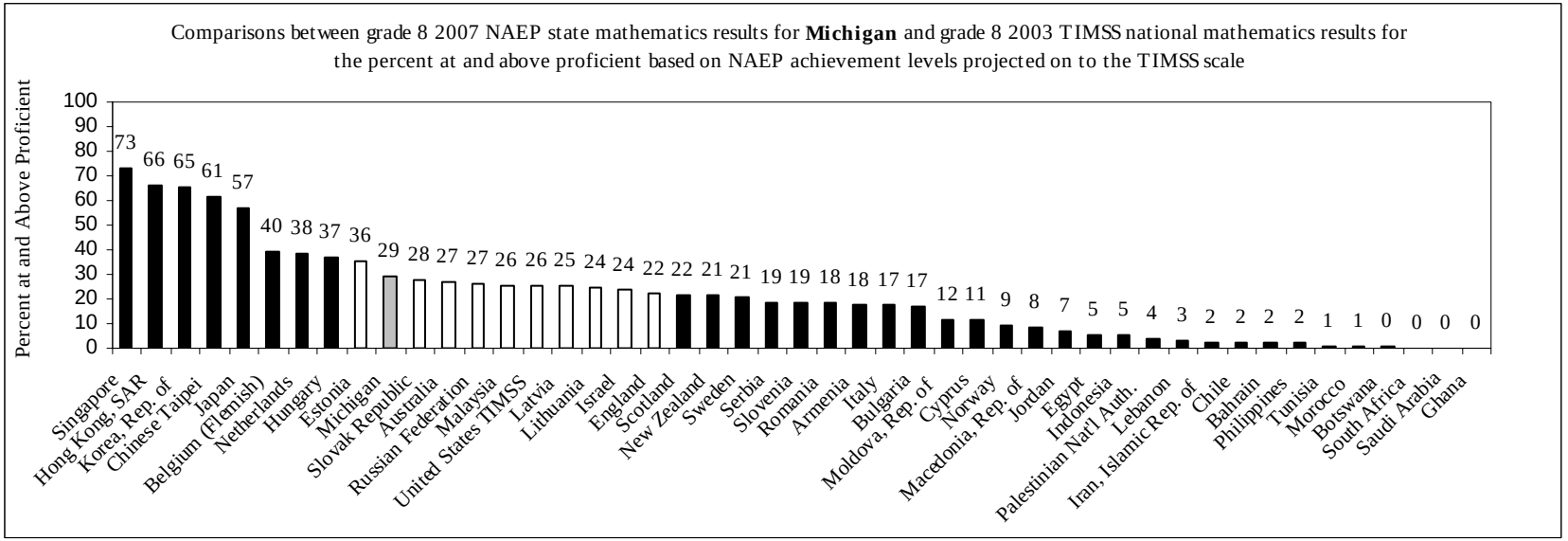
Figure 23: Massachusetts

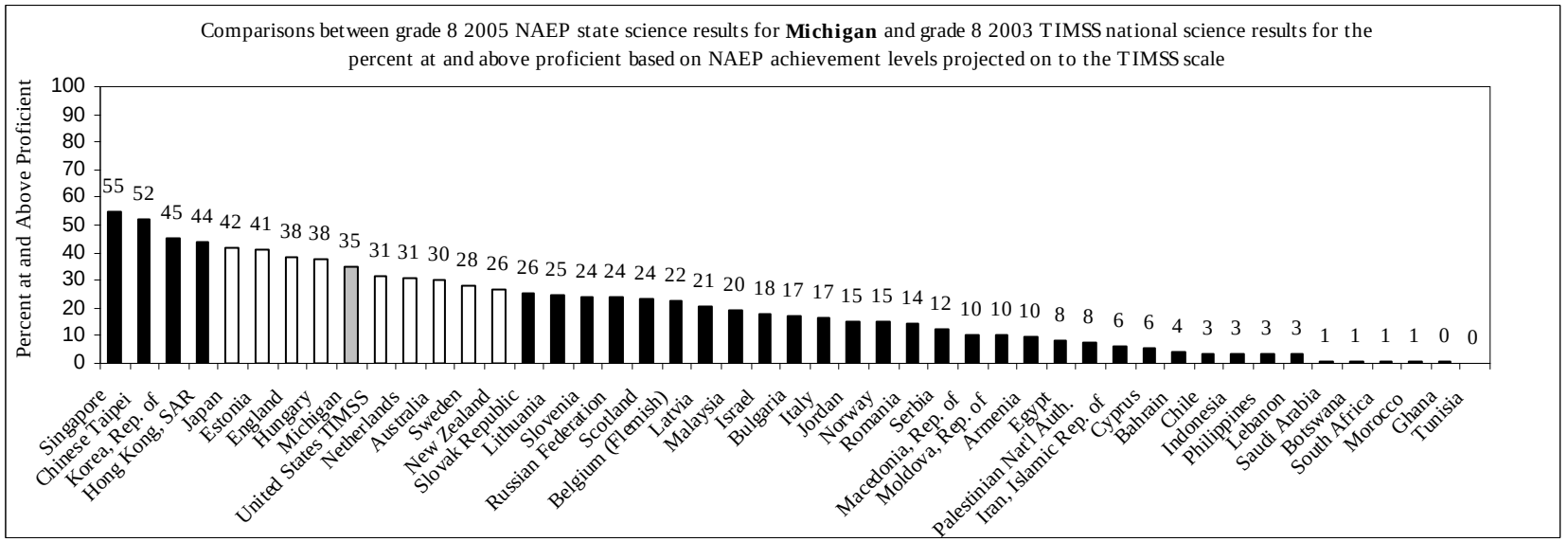




Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

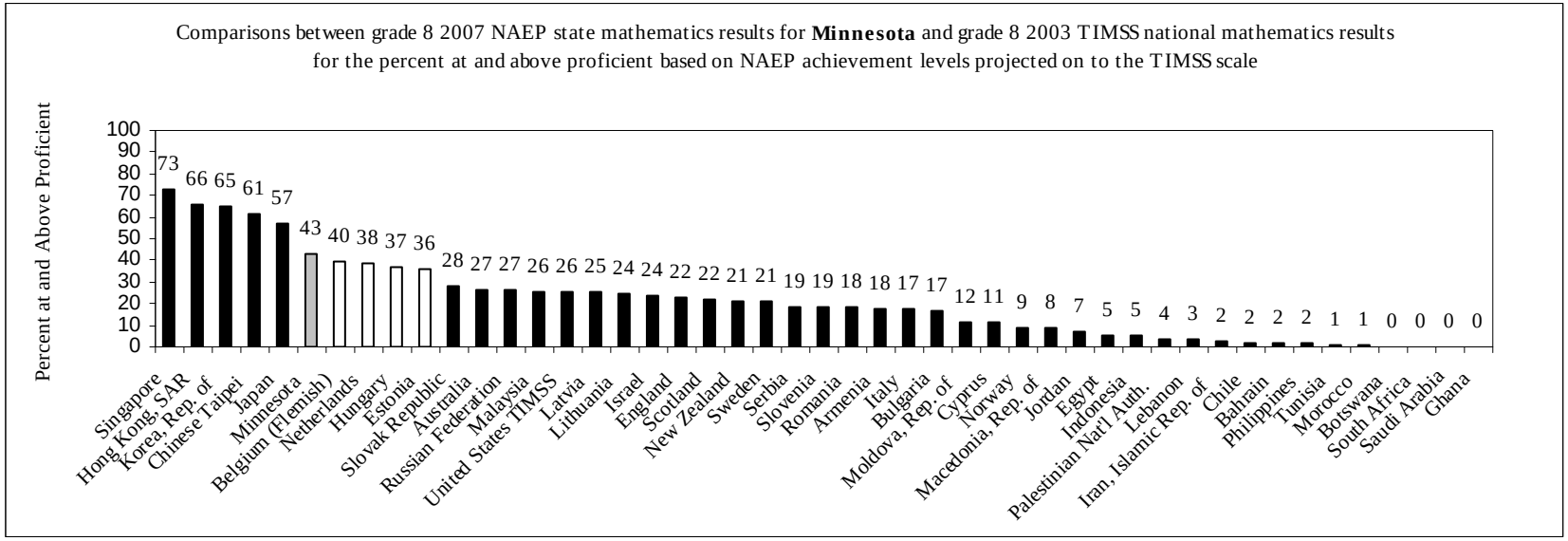
Figure 24: Michigan

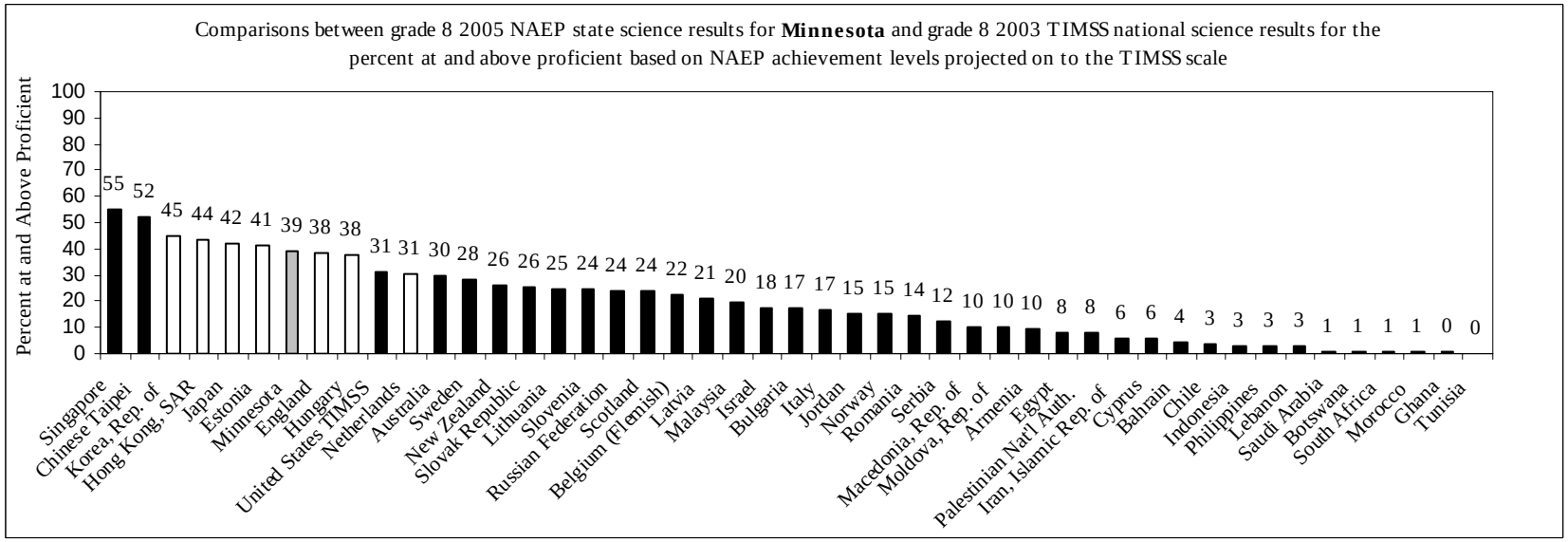




Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

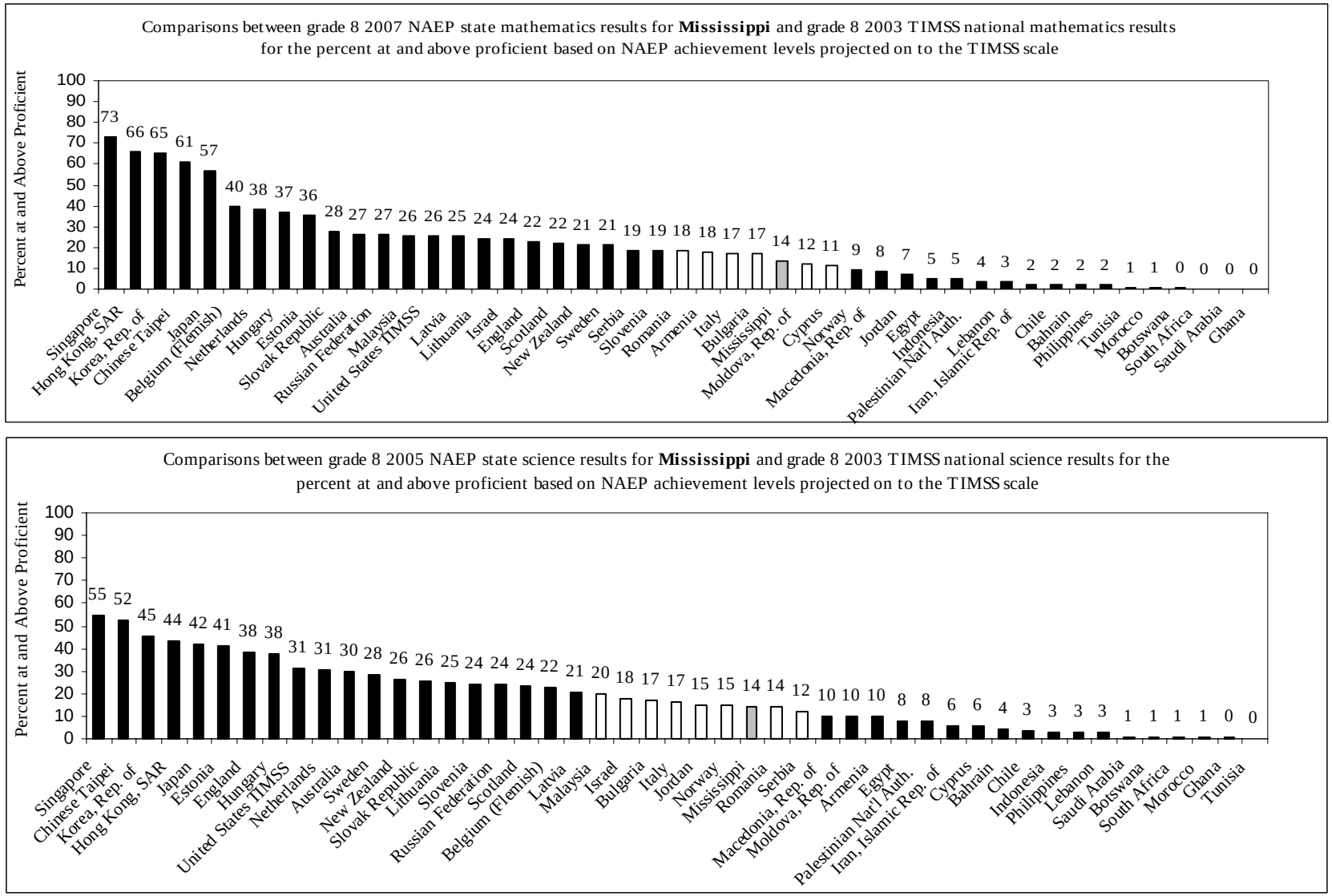
Figure 25: Minnesota





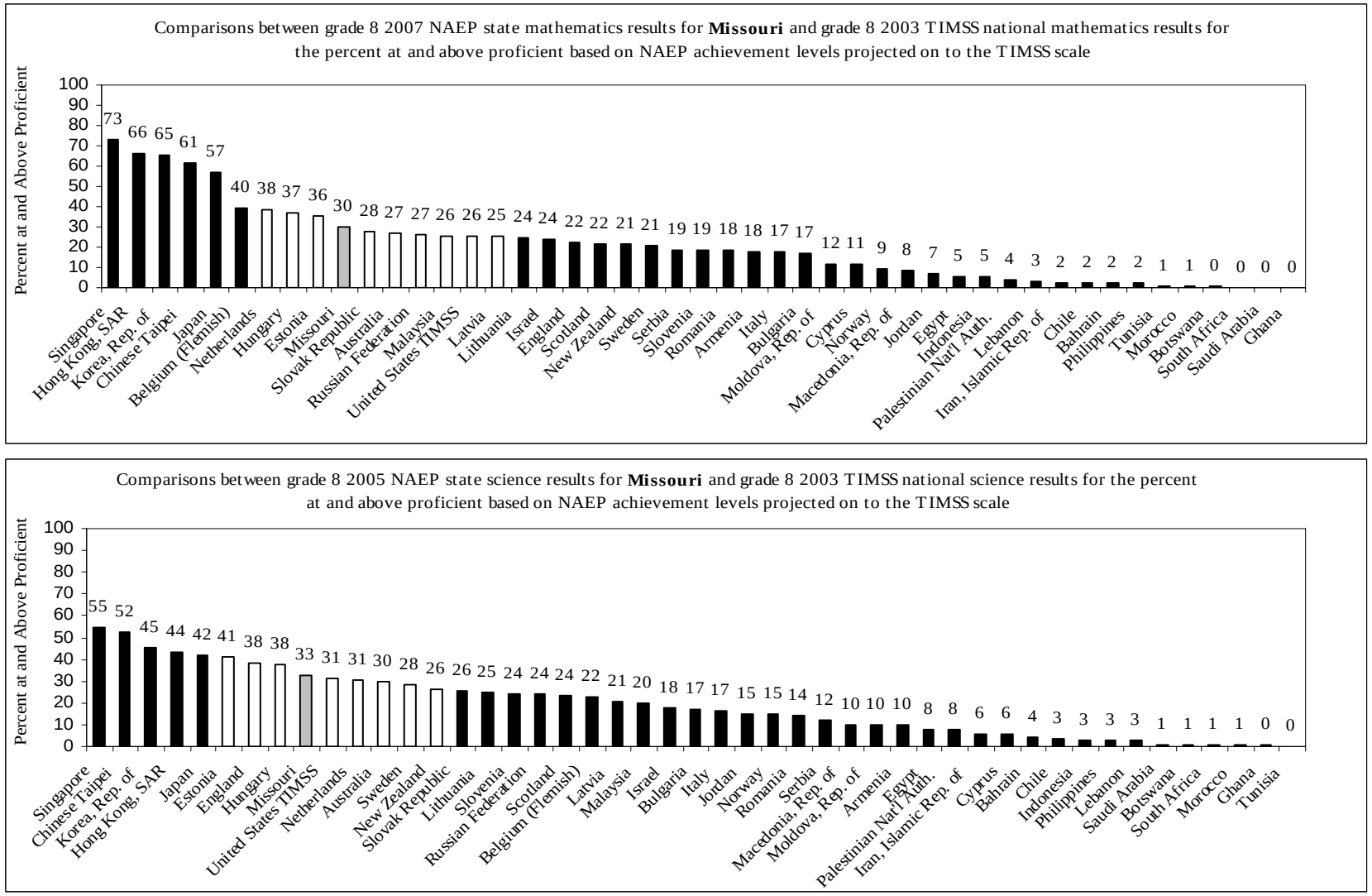
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

Figure 26: Mississippi



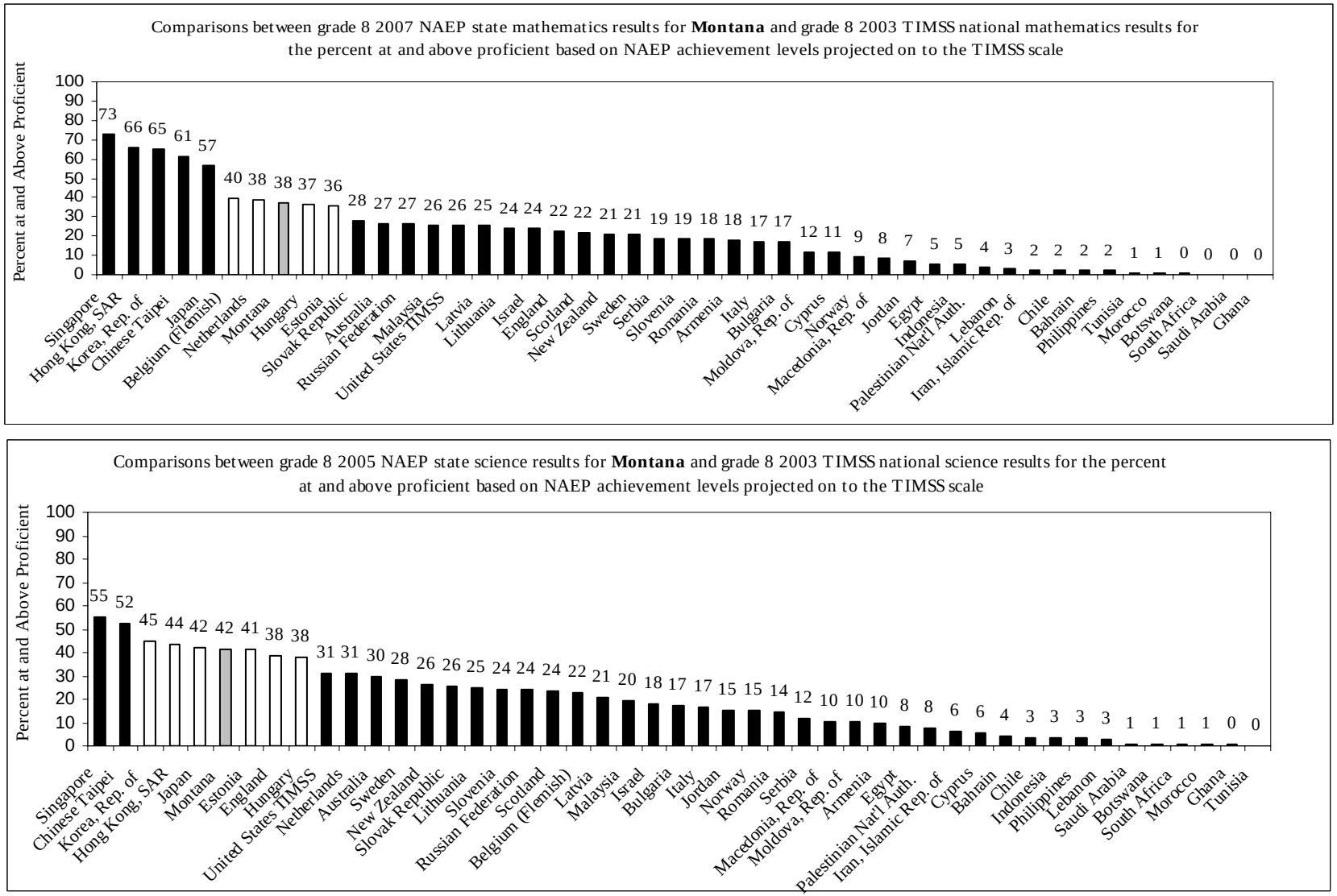
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

Figure 27: Missouri



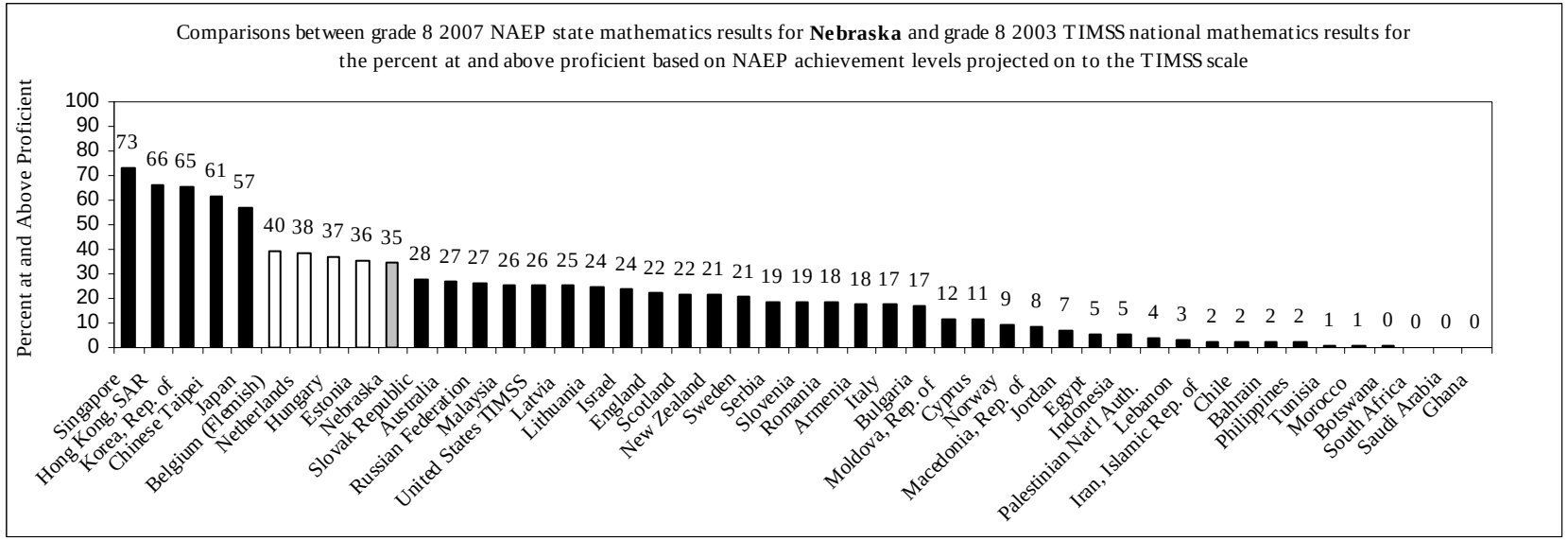
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

Figure 28: Montana



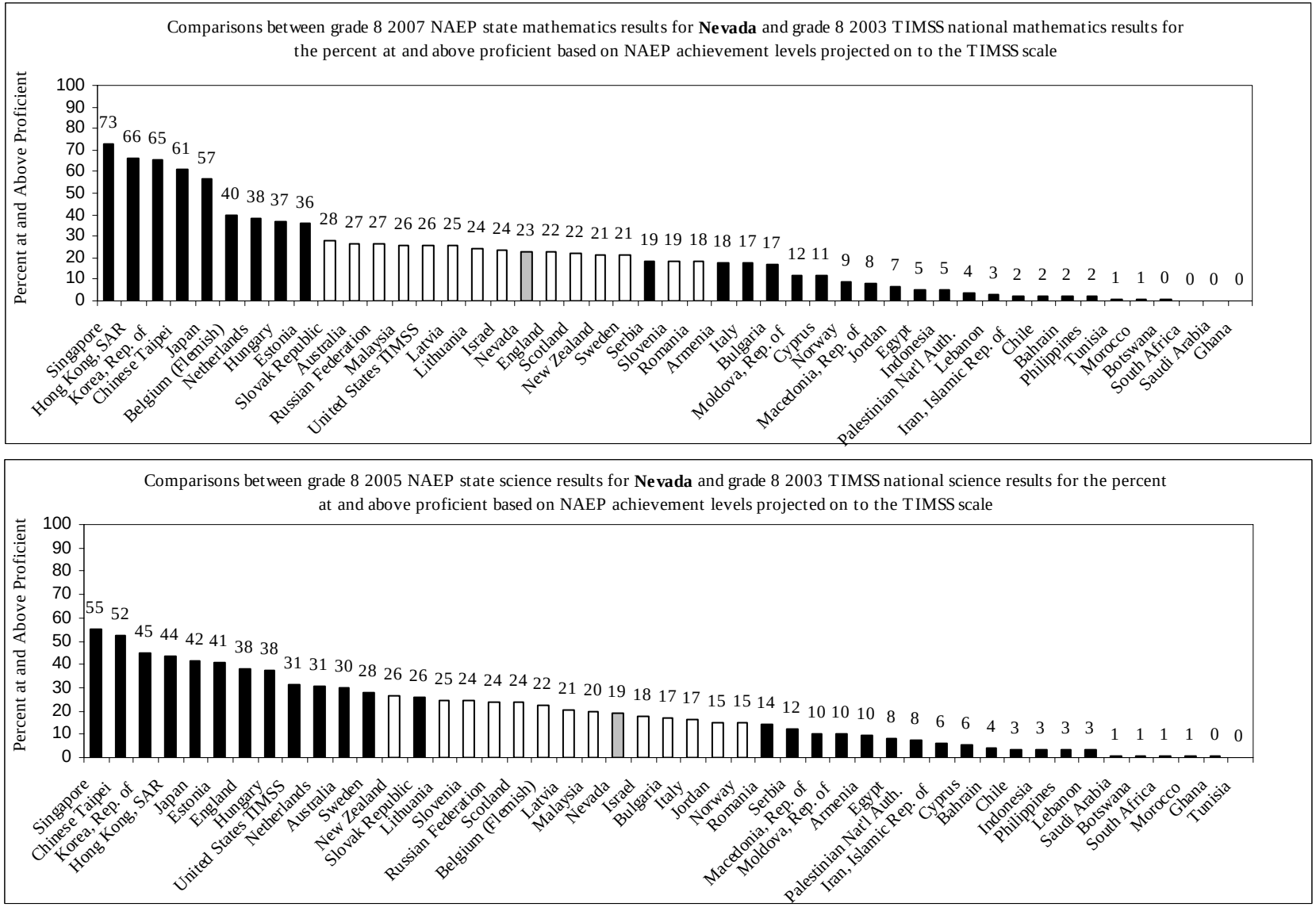
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

Figure 29: Nebraska



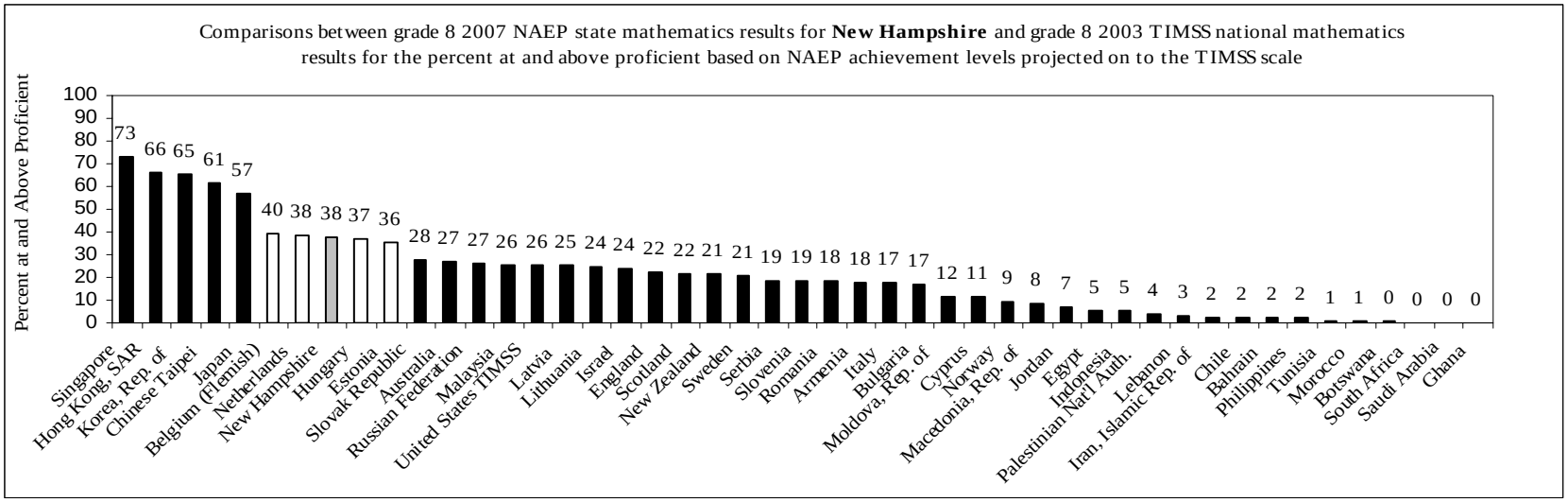
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.
Nebraska did not participate in the grade 8 2005 state NAEP in science.

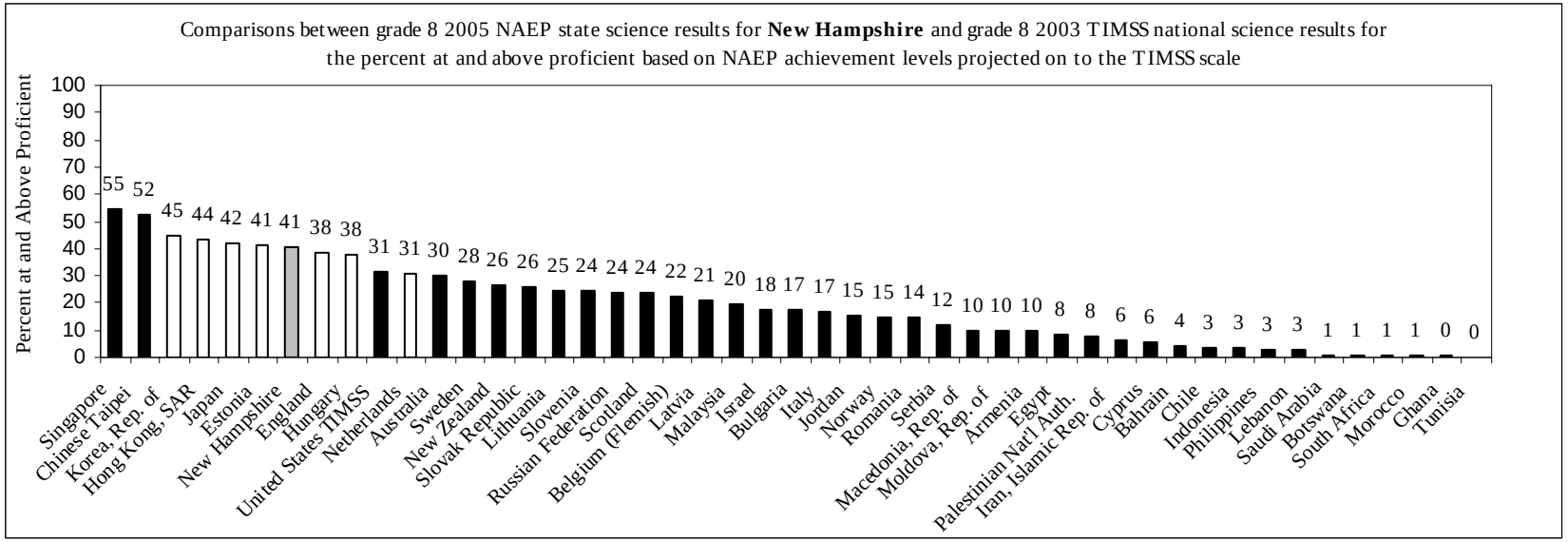
Figure 30: Nevada



Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

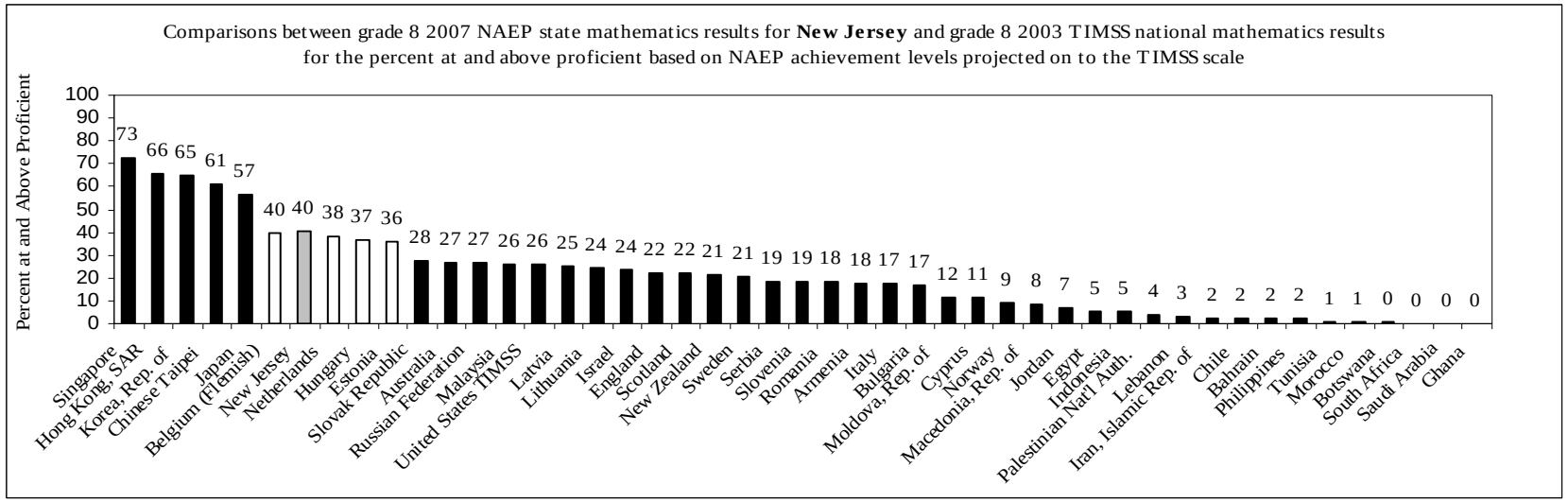
Figure 31: New Hampshire

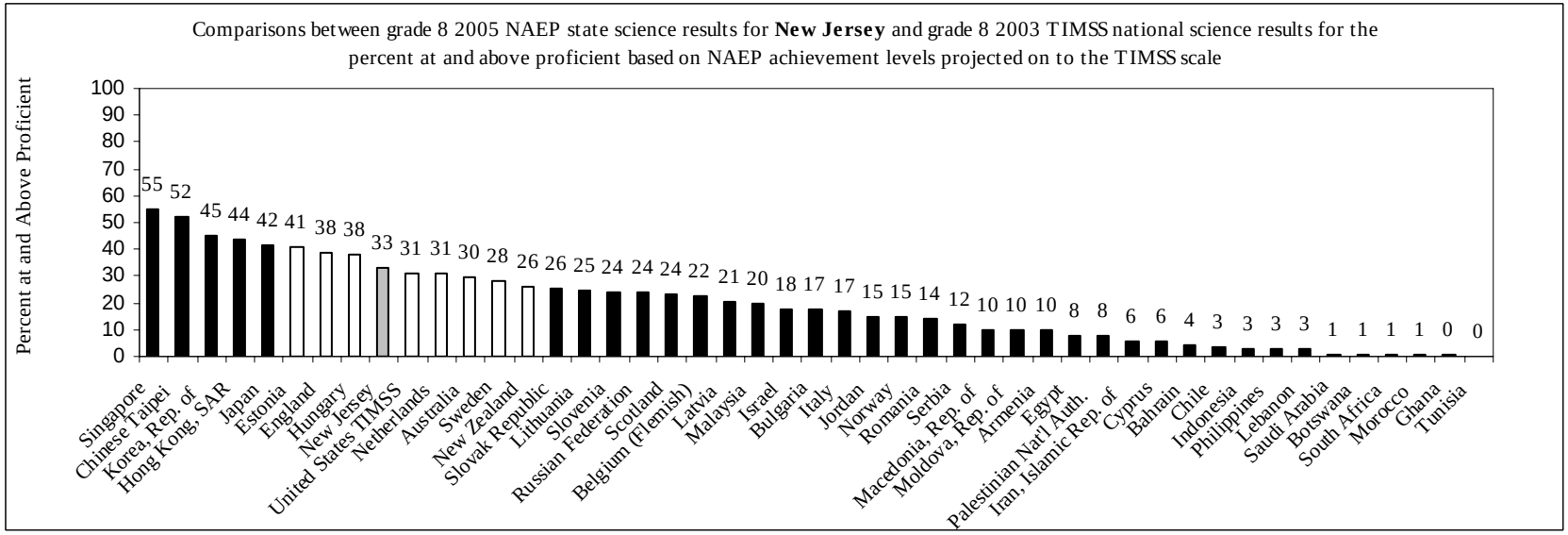




Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

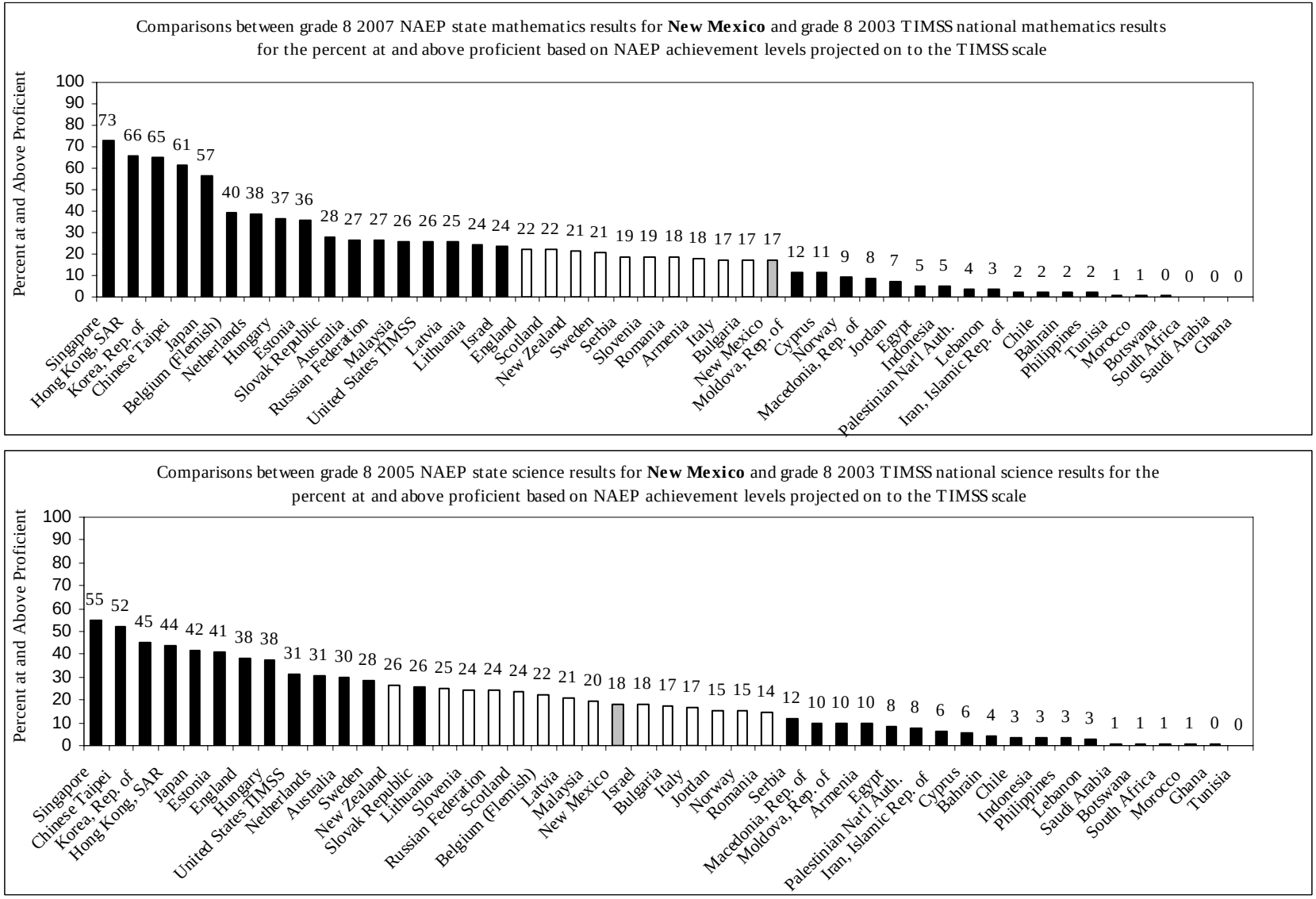
Figure 32: New Jersey





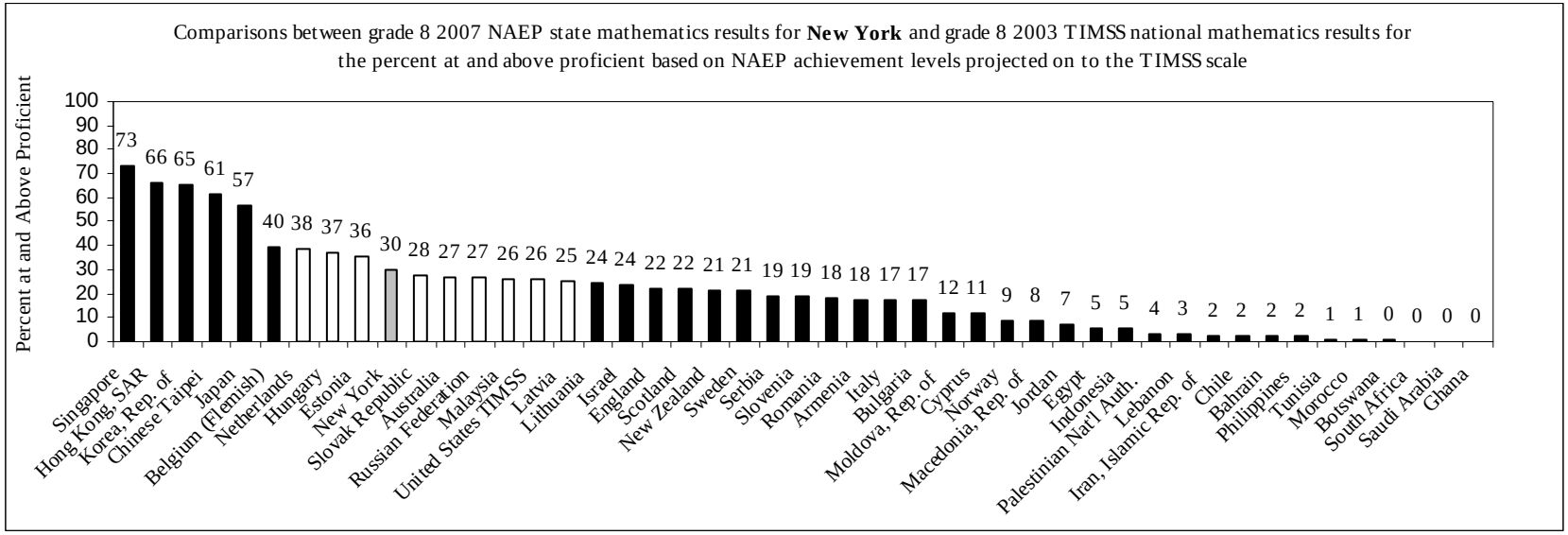
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

Figure 33: New Mexico



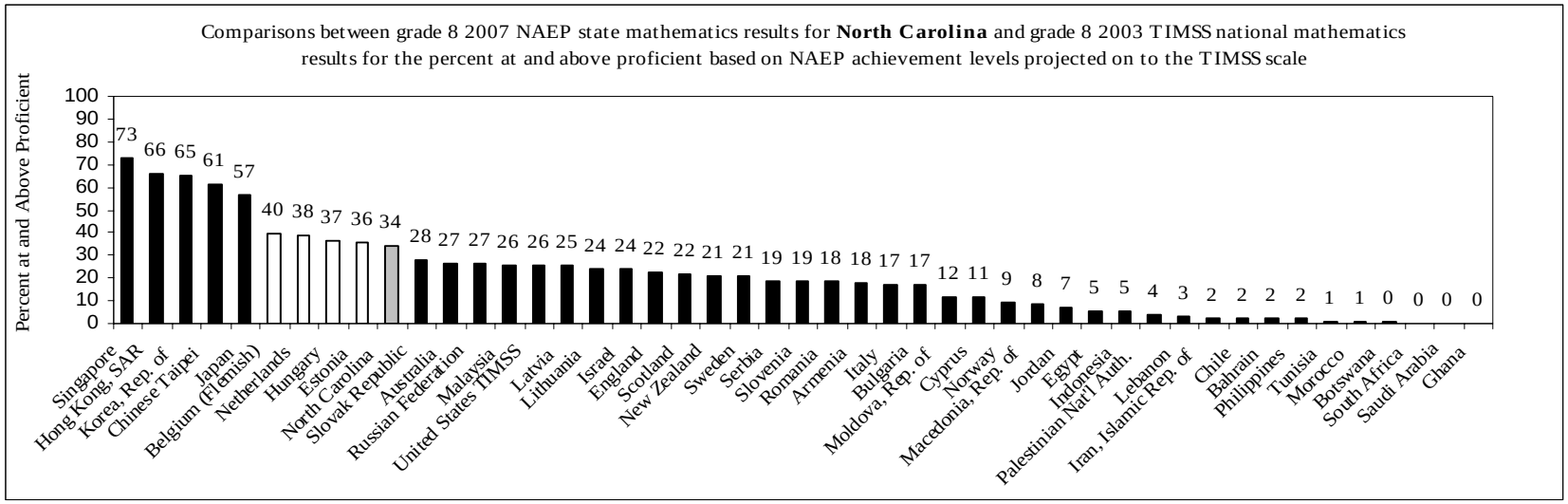
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

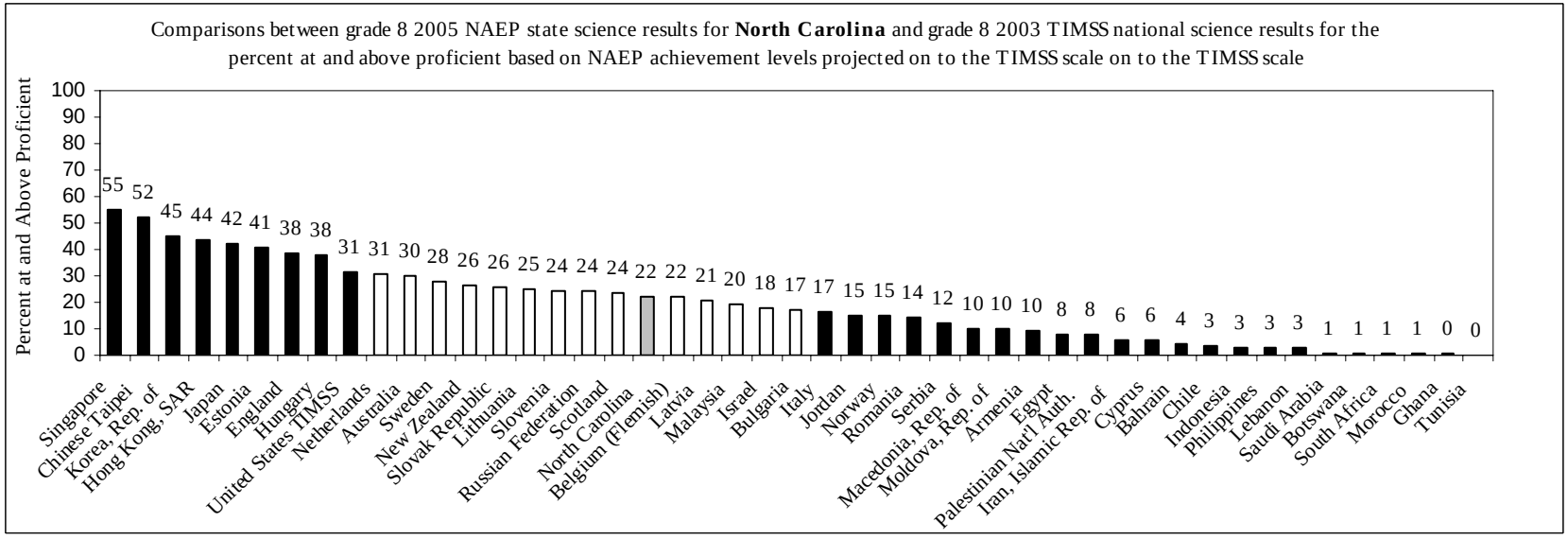
Figure 34: New York



Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.
New York did not participate in the grade 8 2005 state NAEP in science.

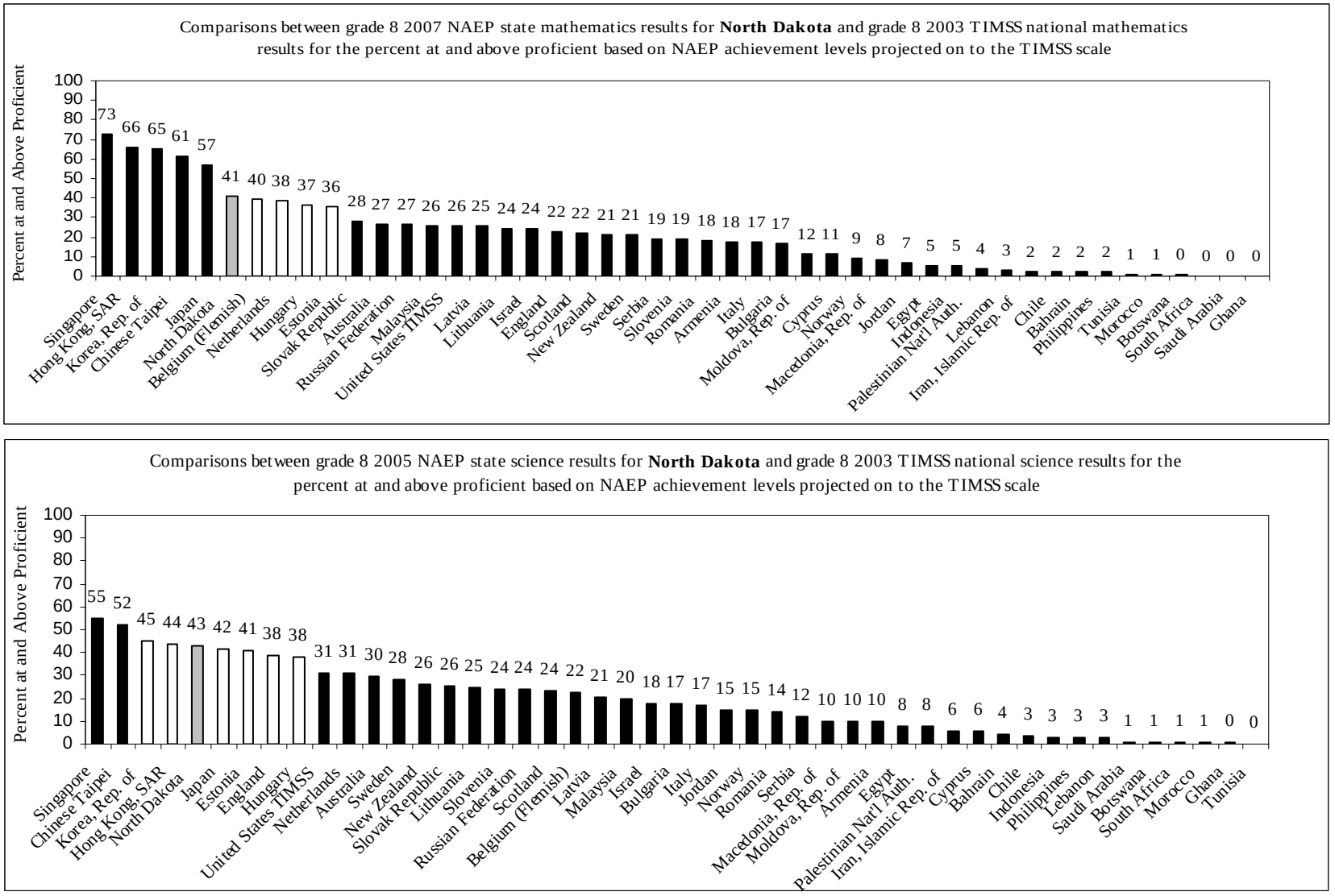
Figure 35: North Carolina





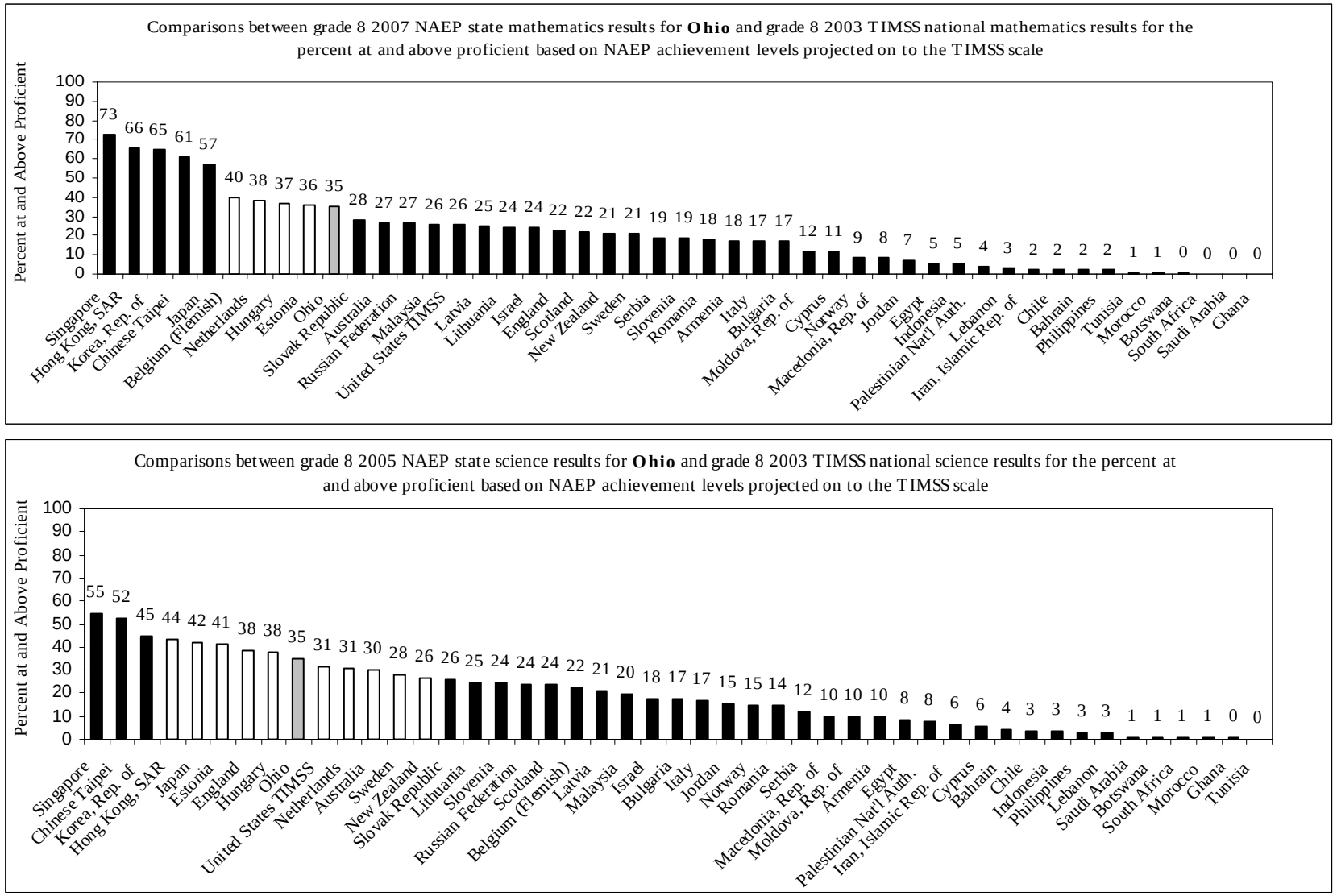
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

Figure 36: North Dakota



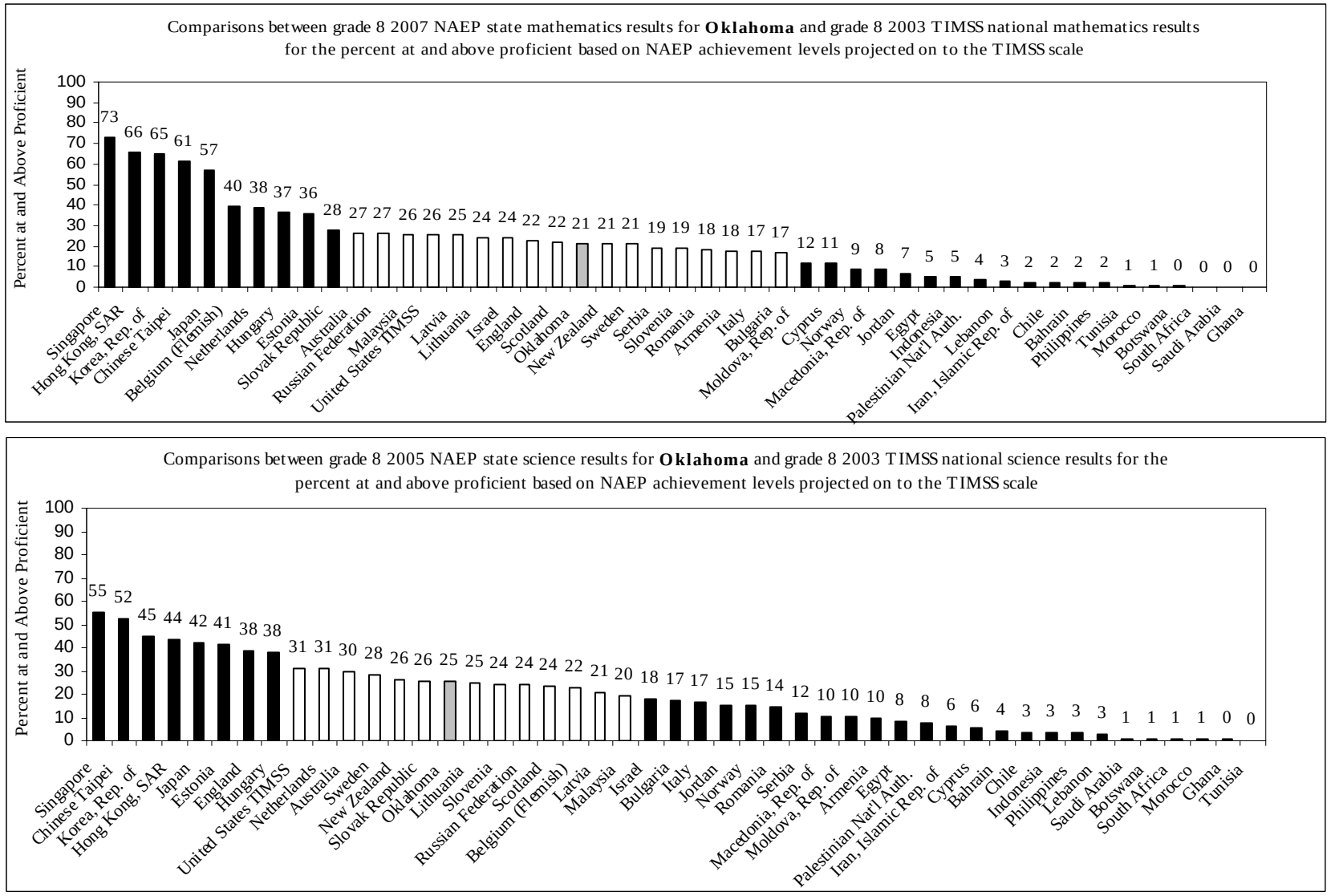
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

Figure 37: Ohio



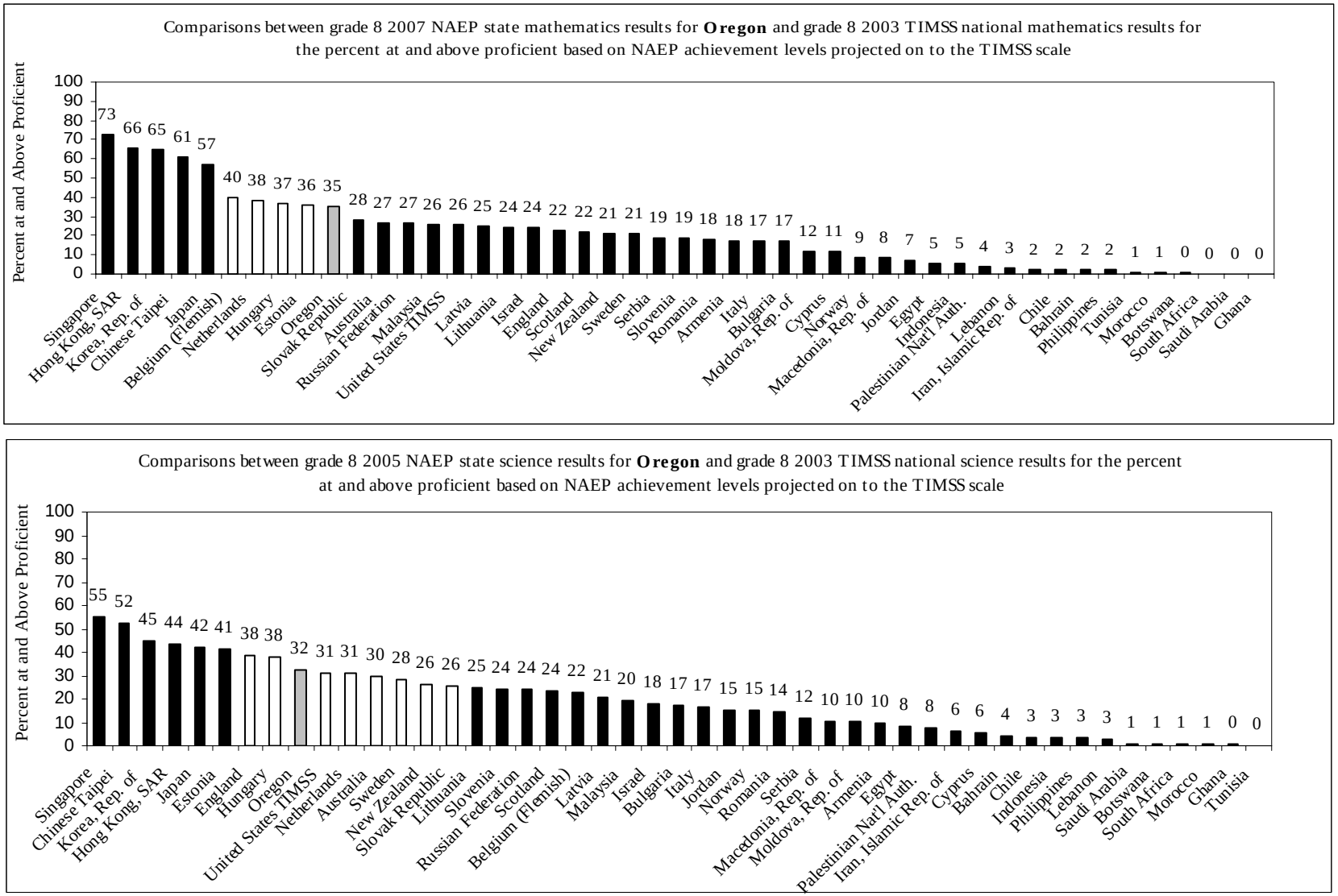
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

Figure 38: Oklahoma



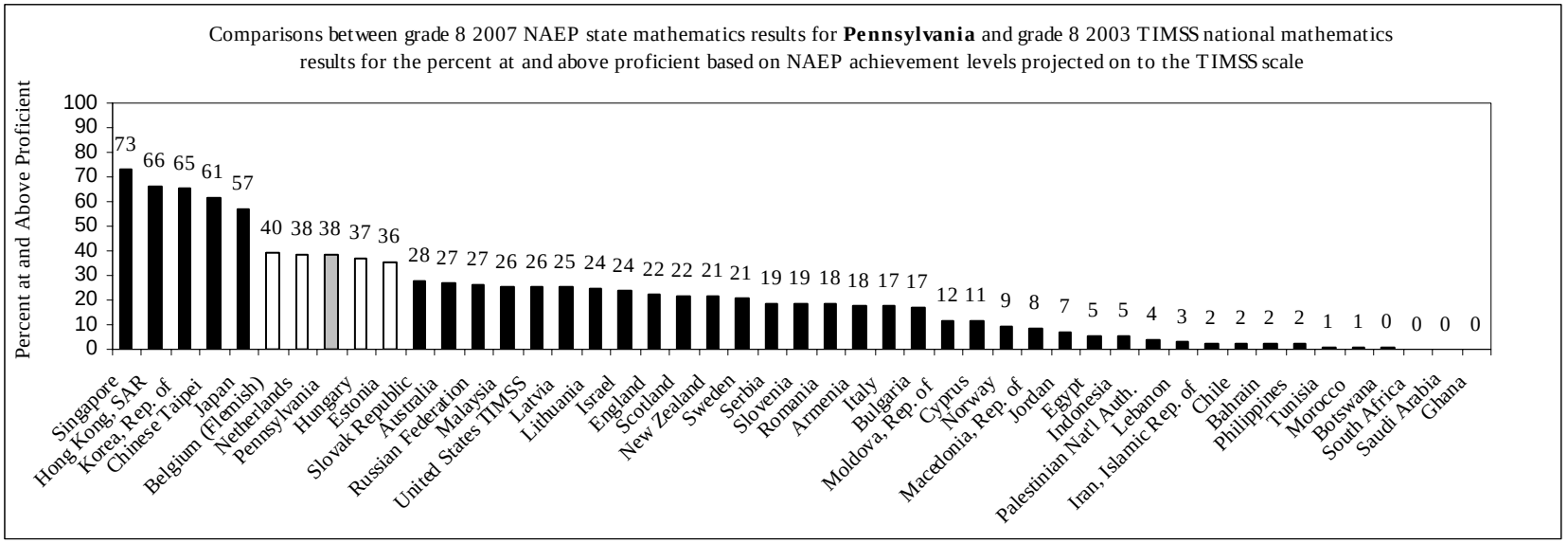
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

Figure 39: Oregon



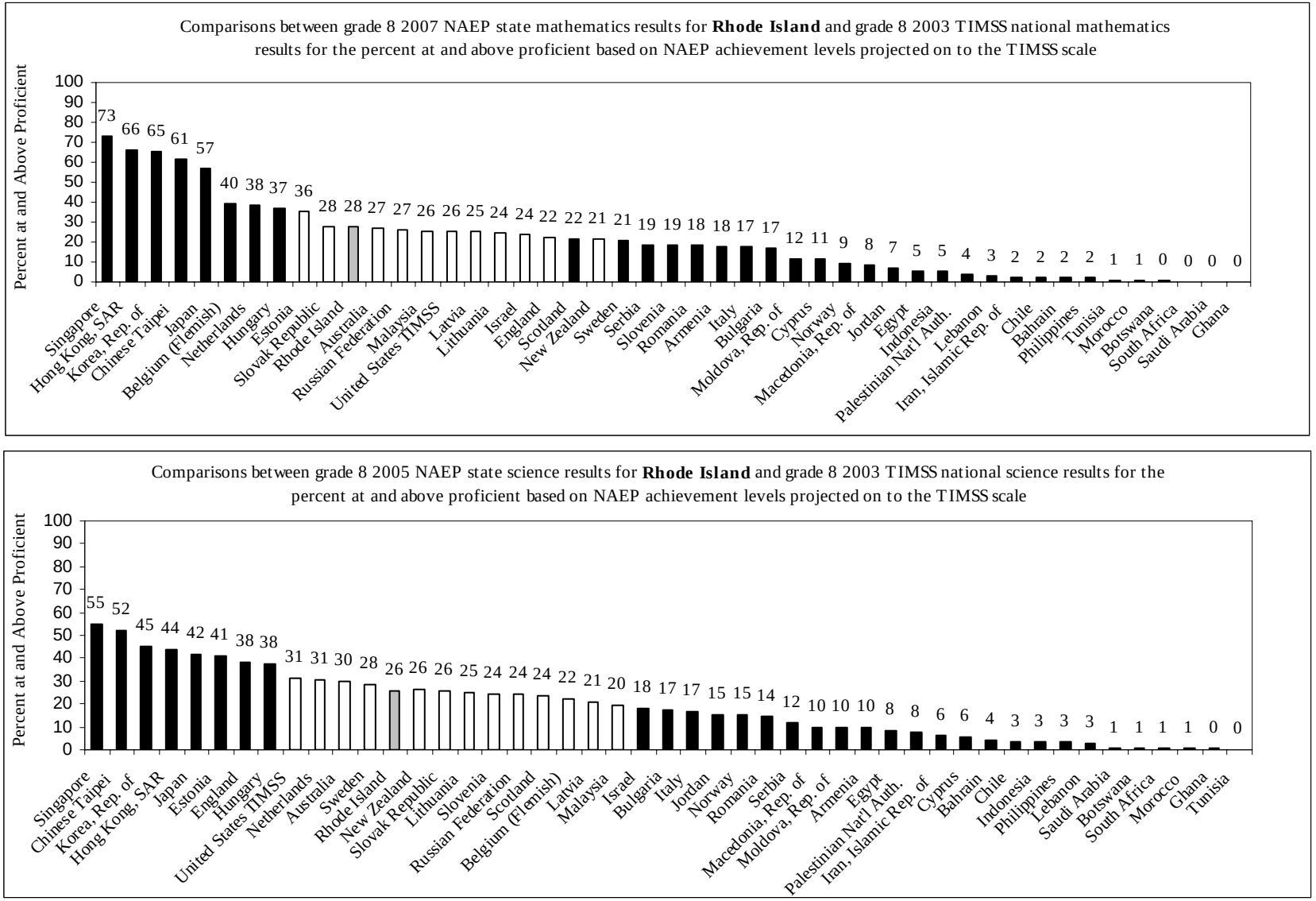
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

Figure40: Pennsylvania



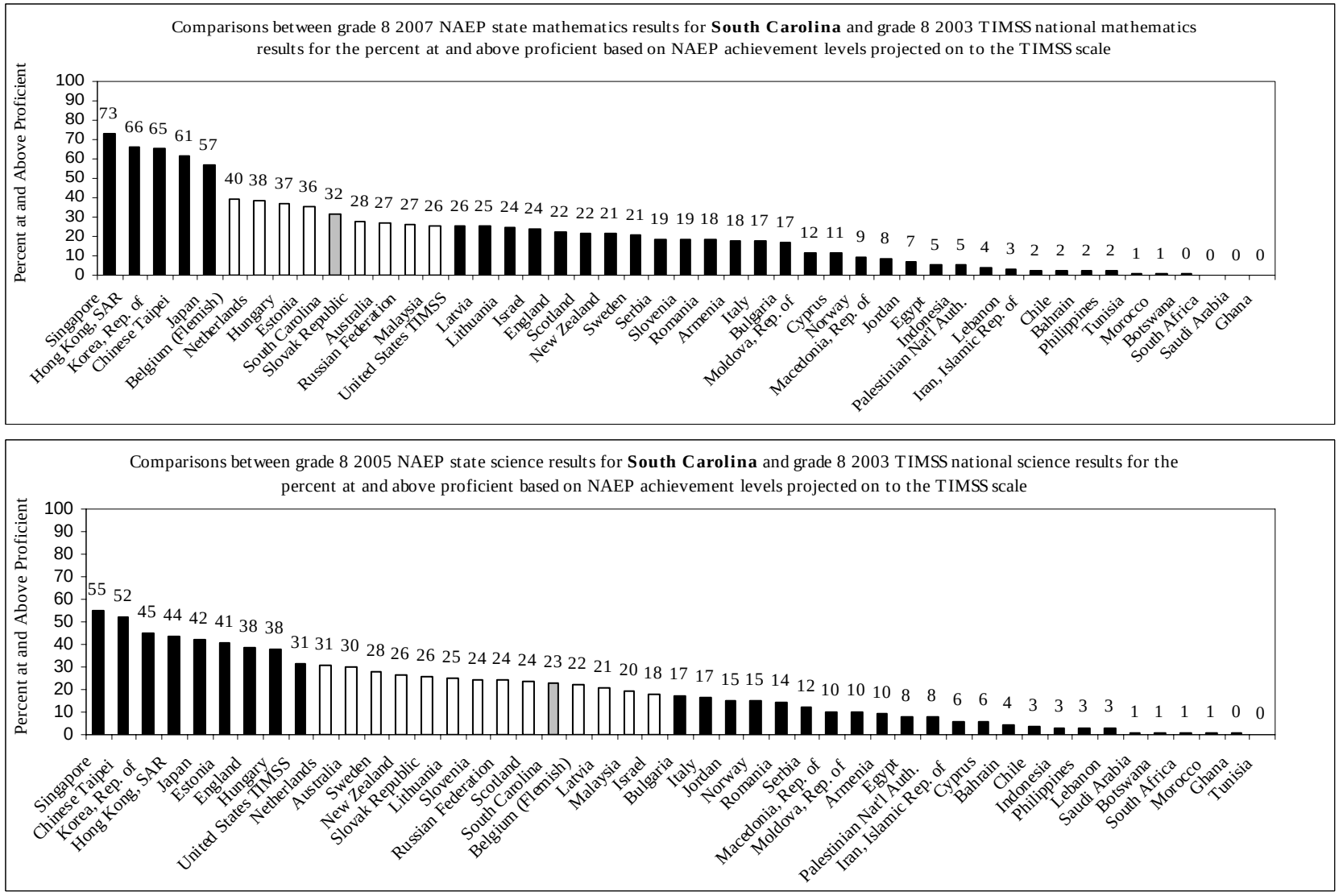
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.
Pennsylvania did not participate in the grade 8 2005 state NAEP in science.

Figure 41: Rhode Island



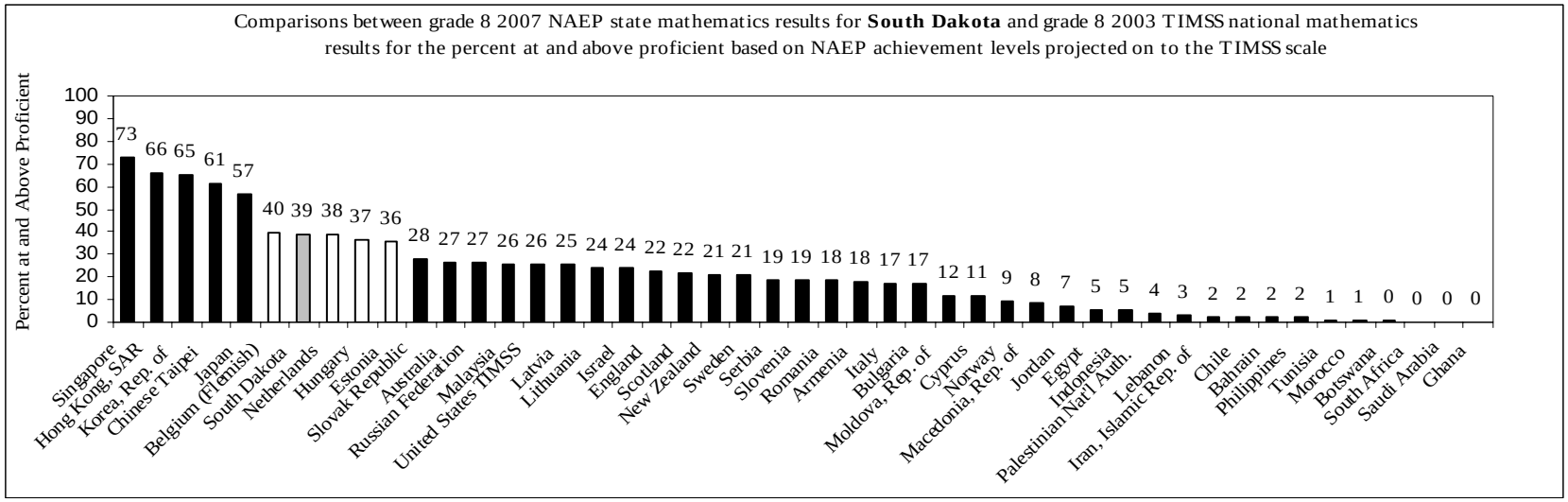
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

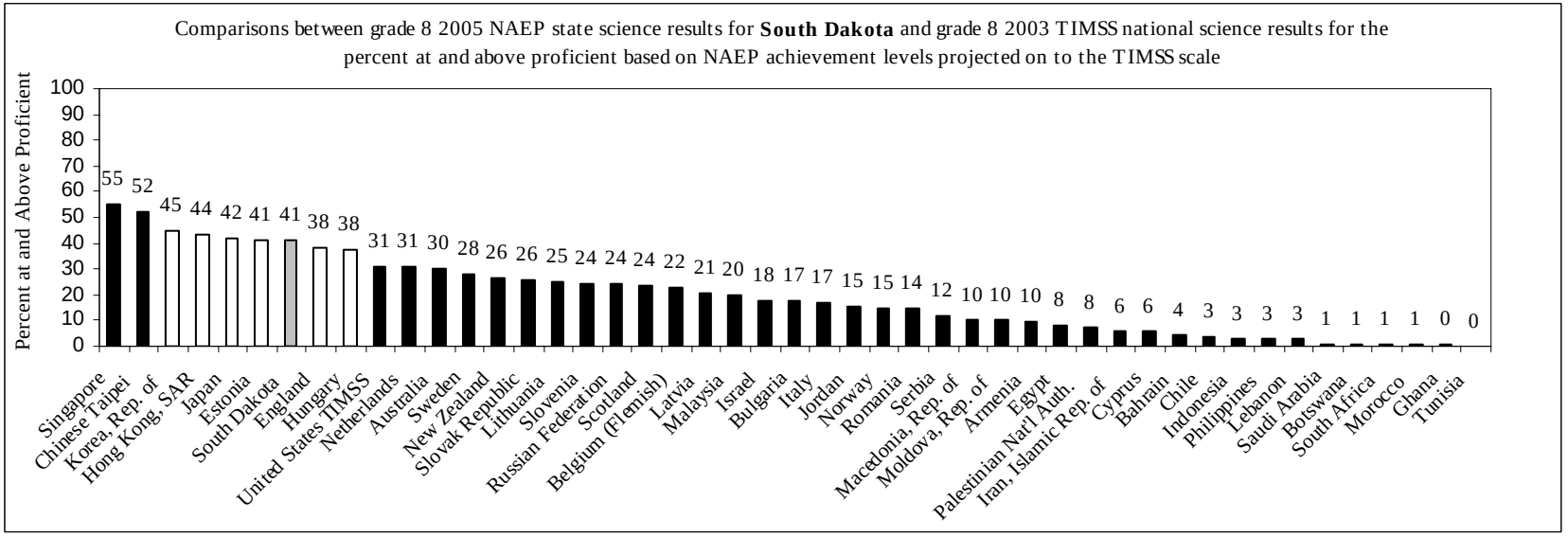
Figure 42: South Carolina



Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

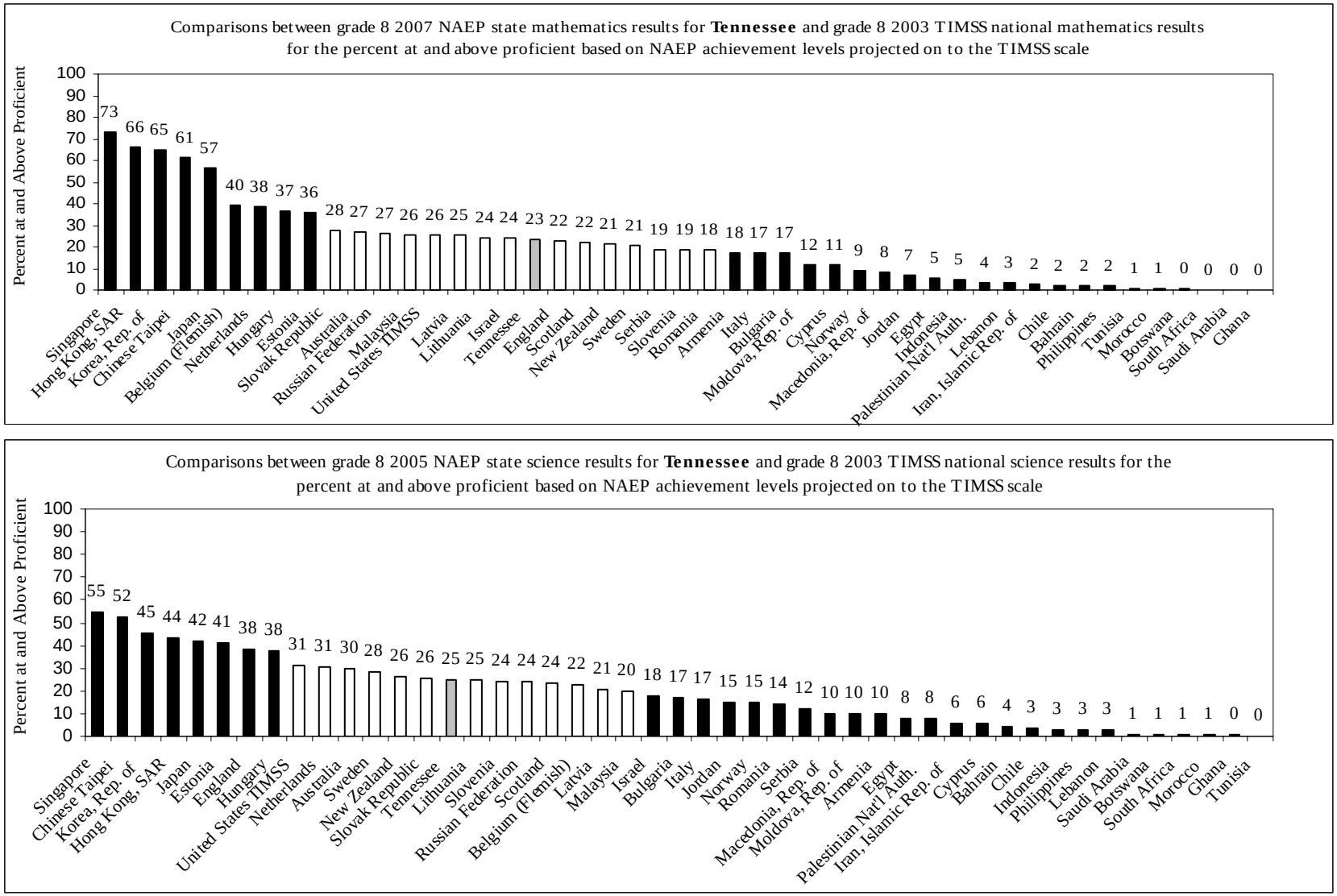
Figure 43: South Dakota





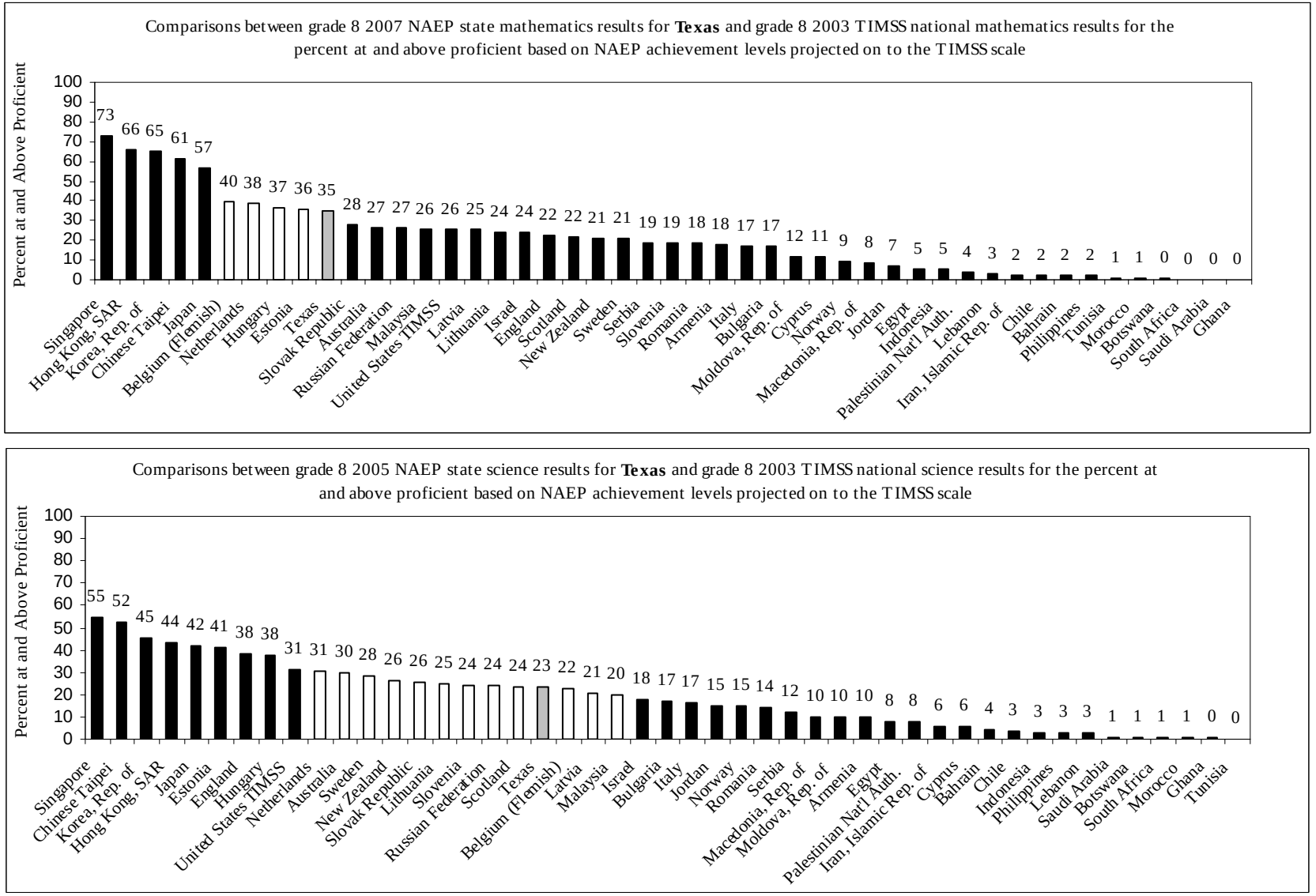
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

Figure 44: Tennessee



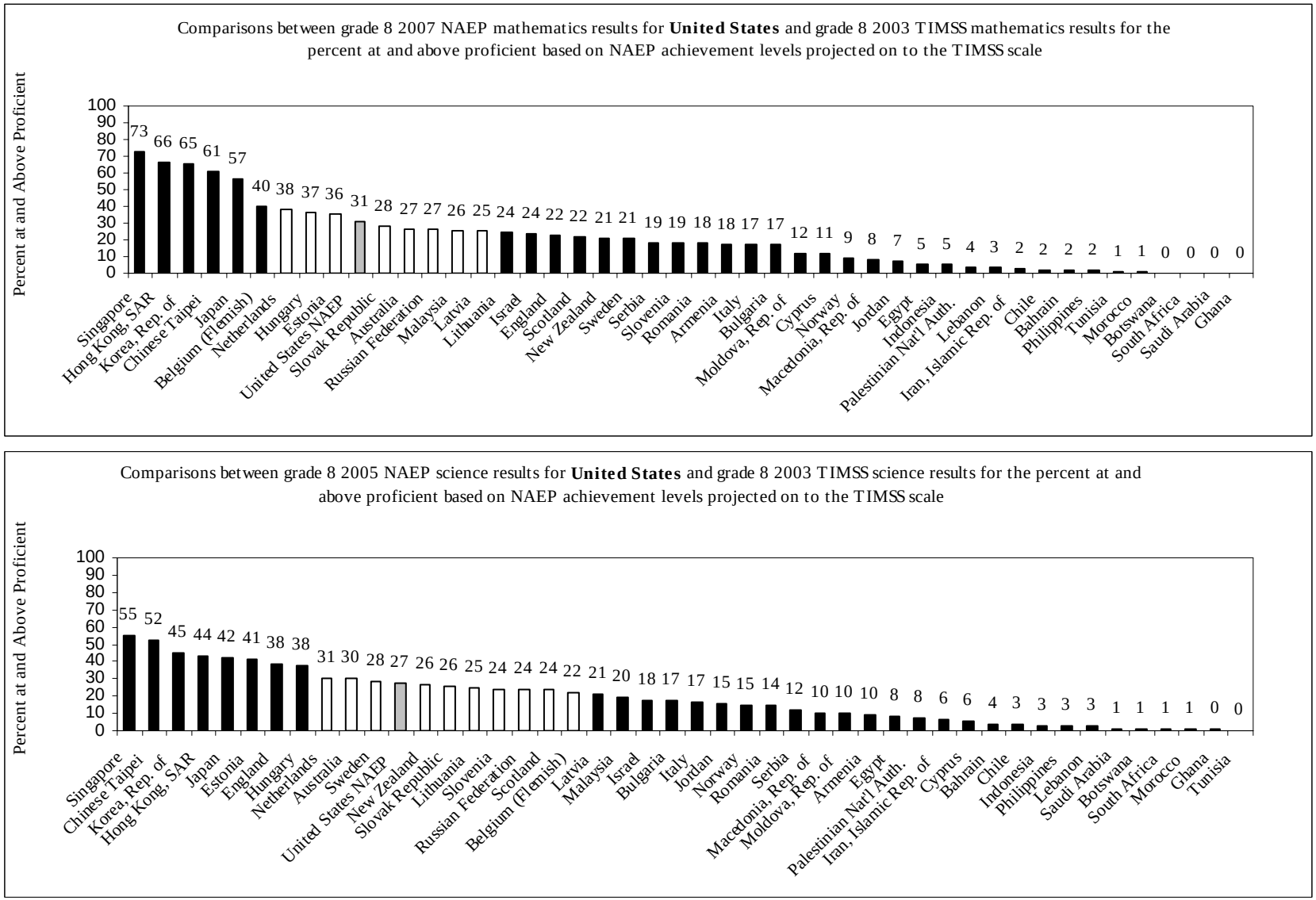
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations,* AIR: Wash., DC, 2007.

Figure 45: Texas



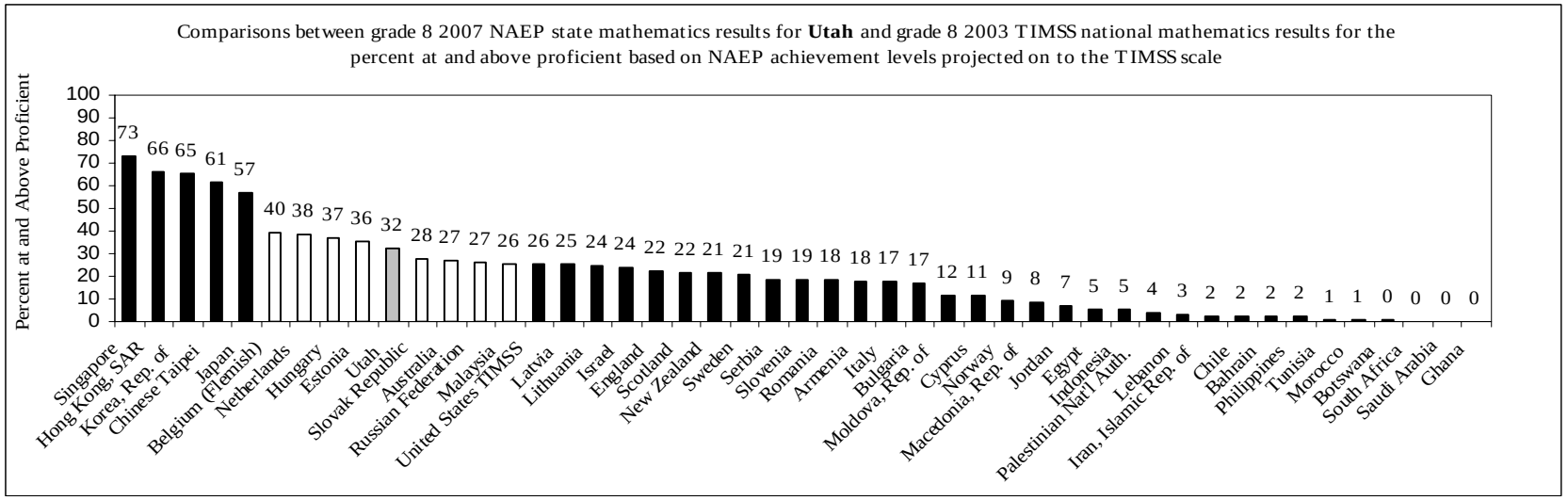
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

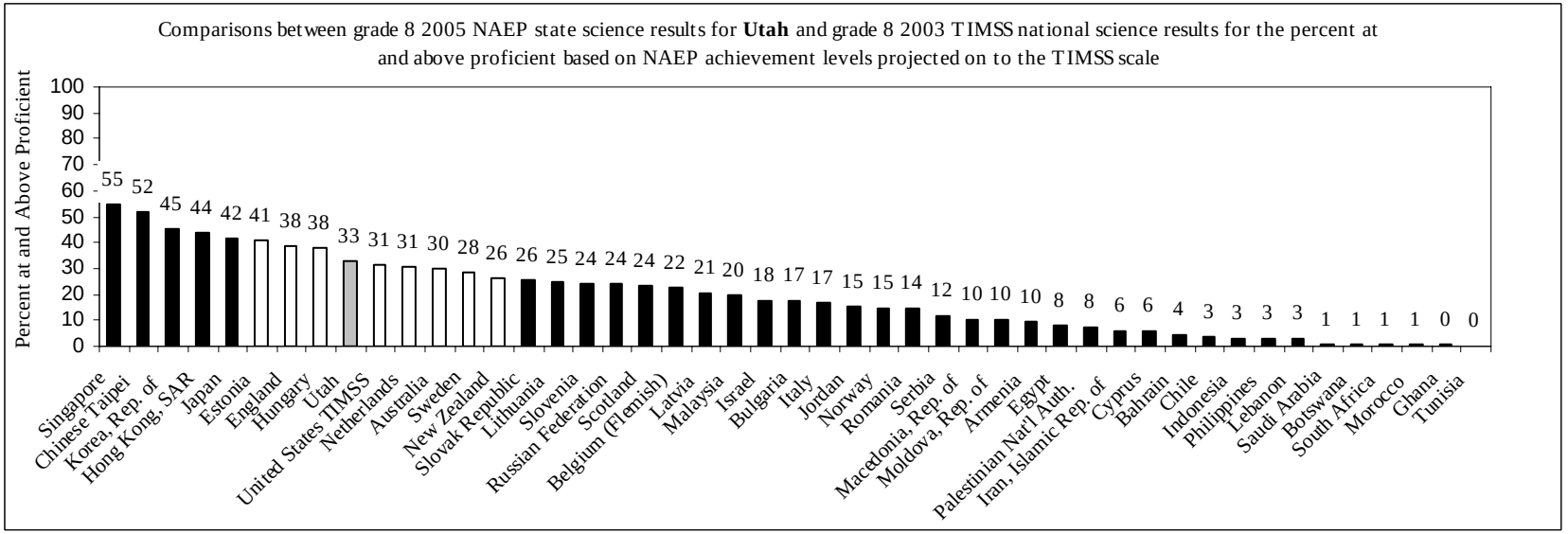
Figure 46: United States



Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

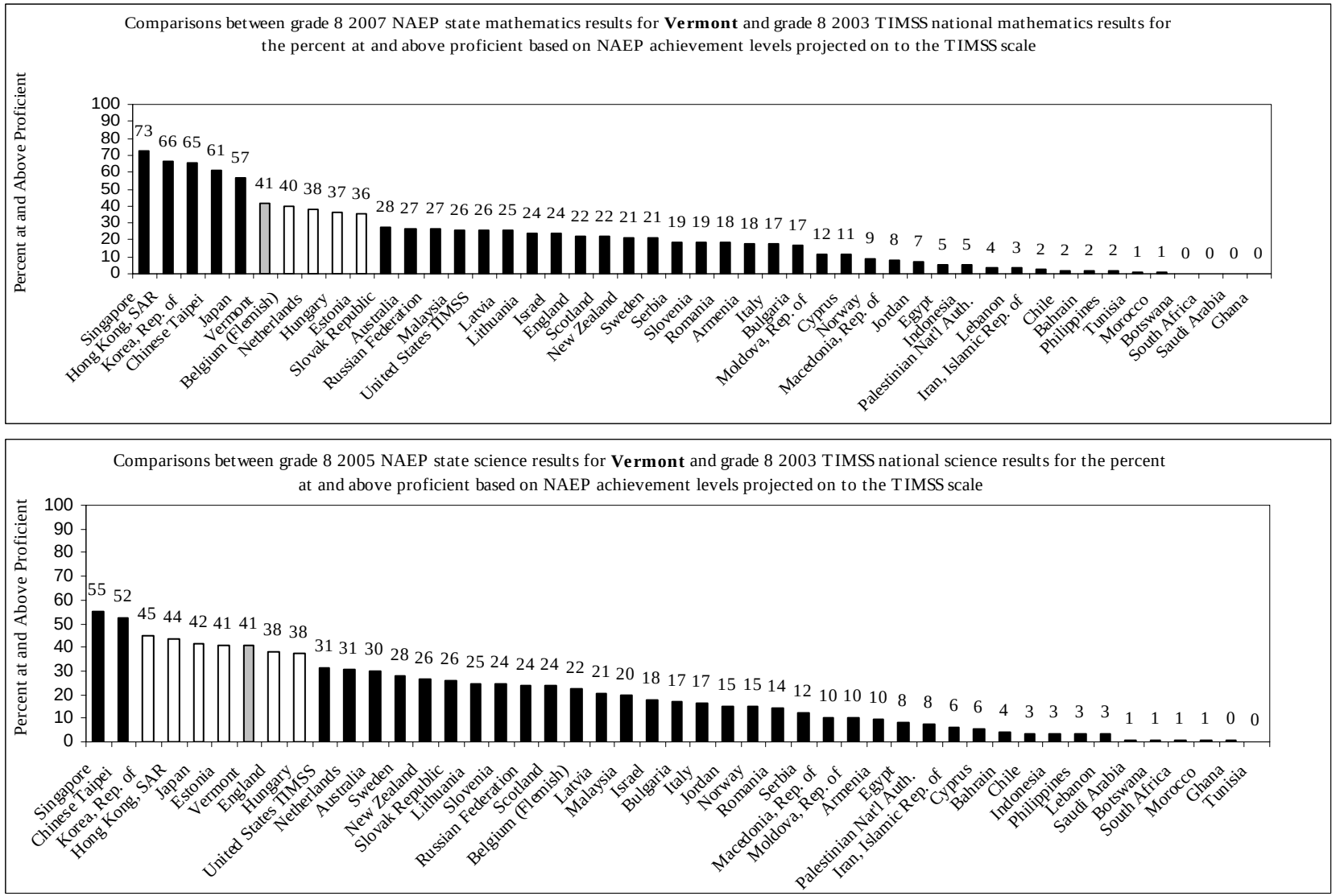
Figure 47: Utah





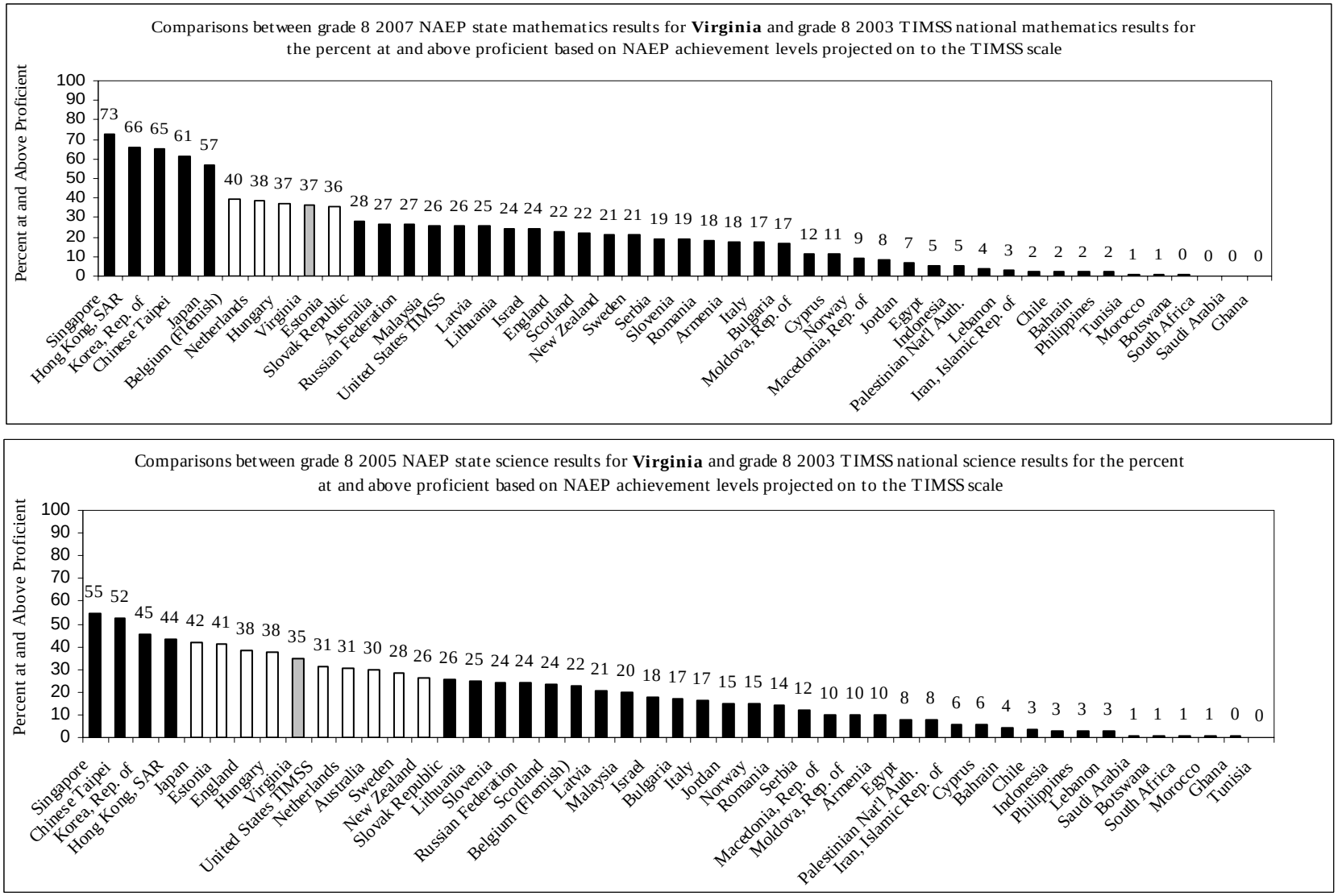
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

Figure 48: Vermont



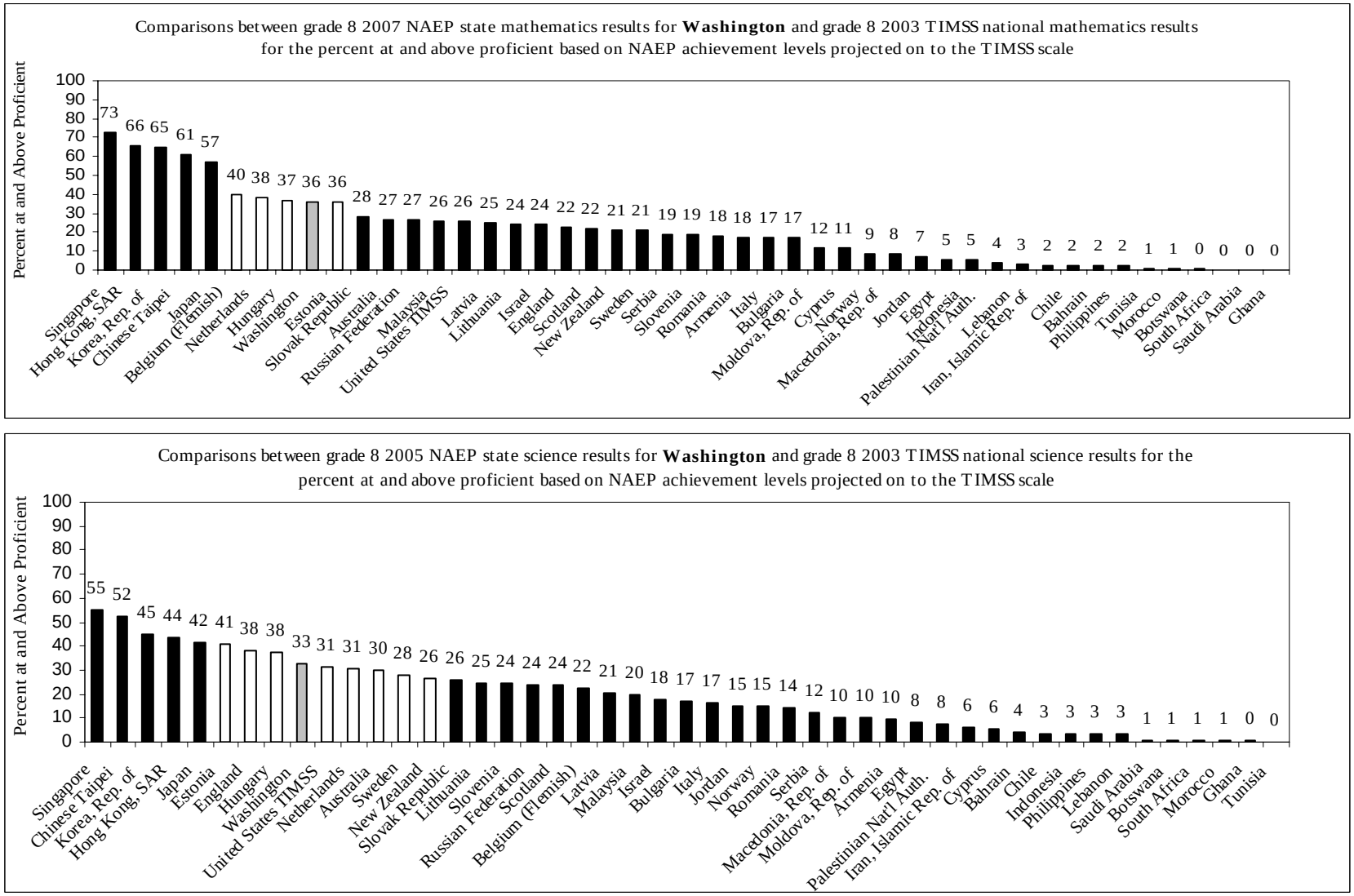
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

Figure 49: Virginia



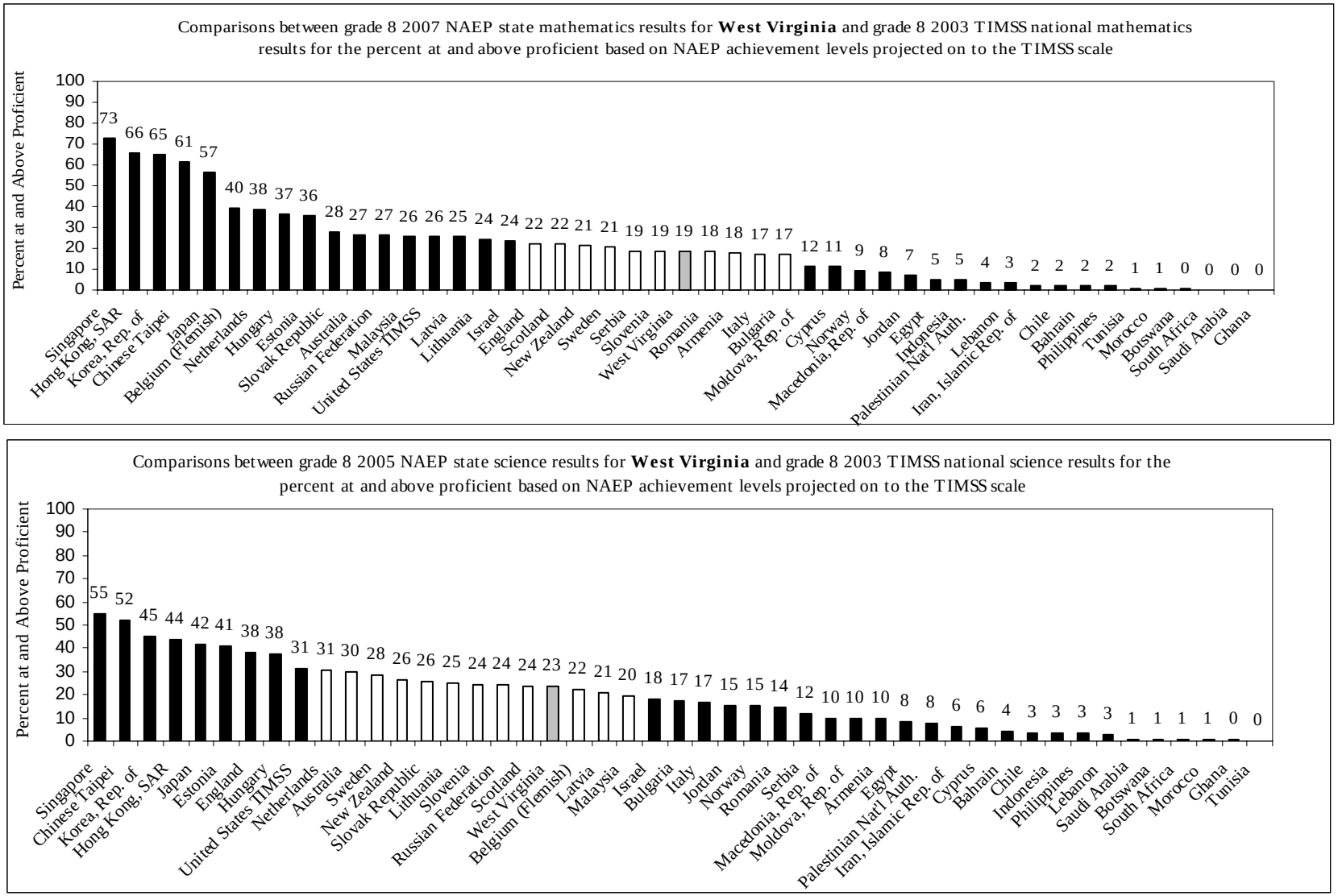
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

Figure 50: Washington



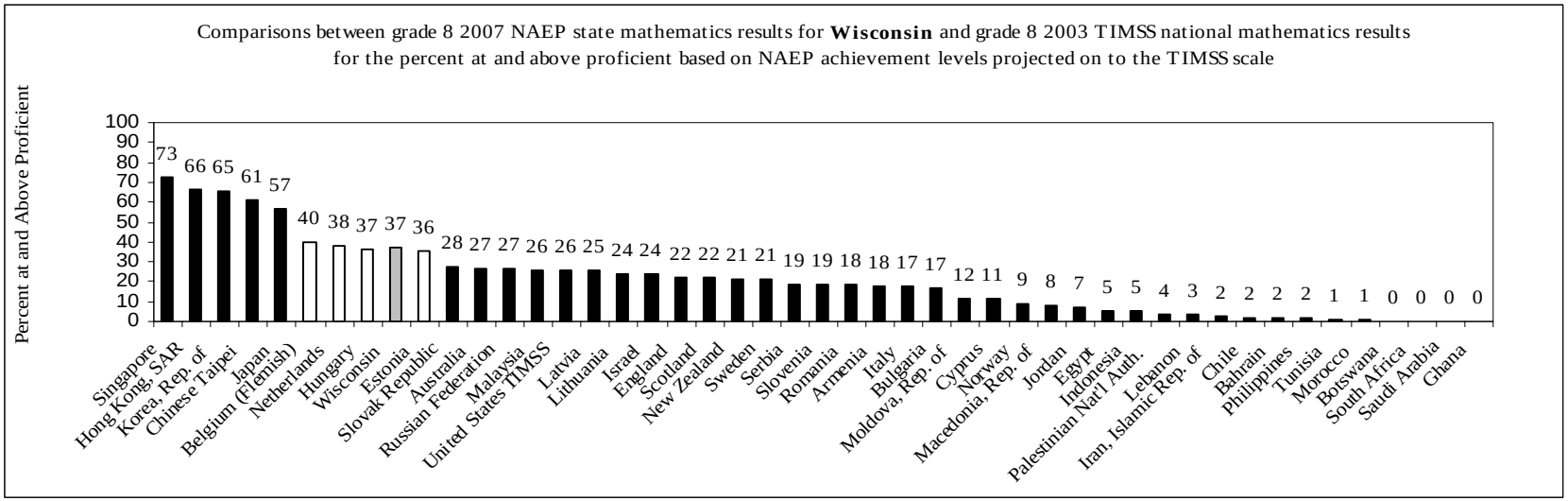
Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

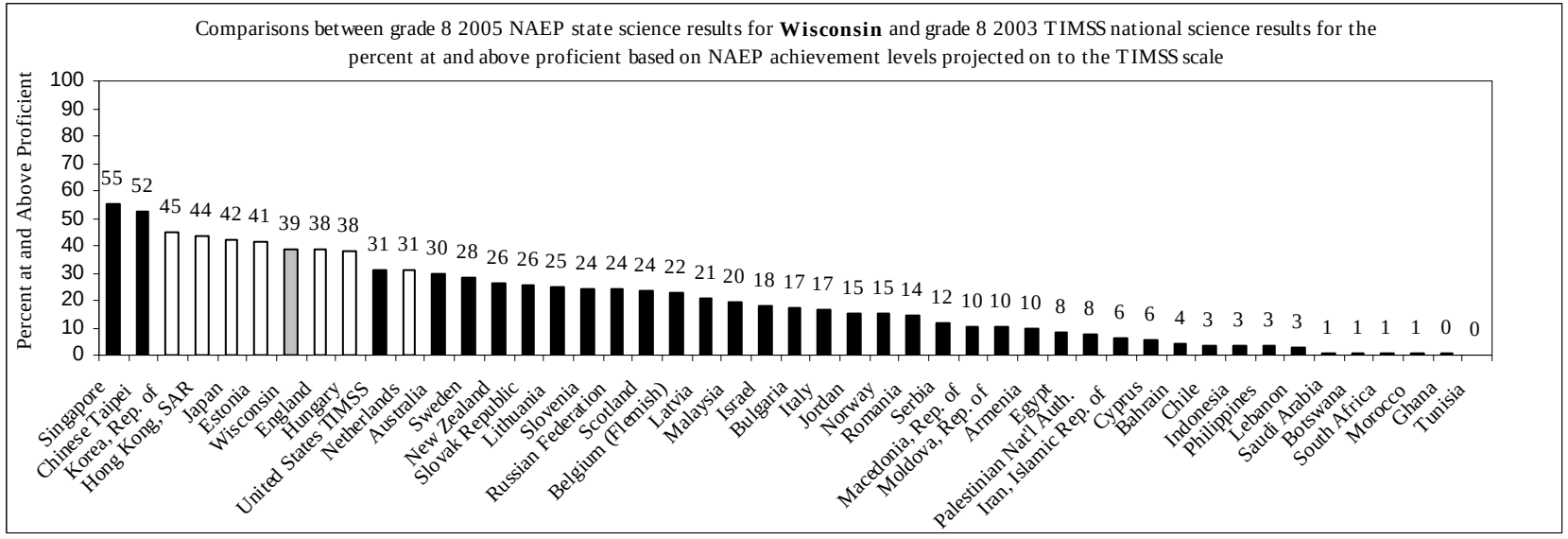
Figure 51: West Virginia



Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

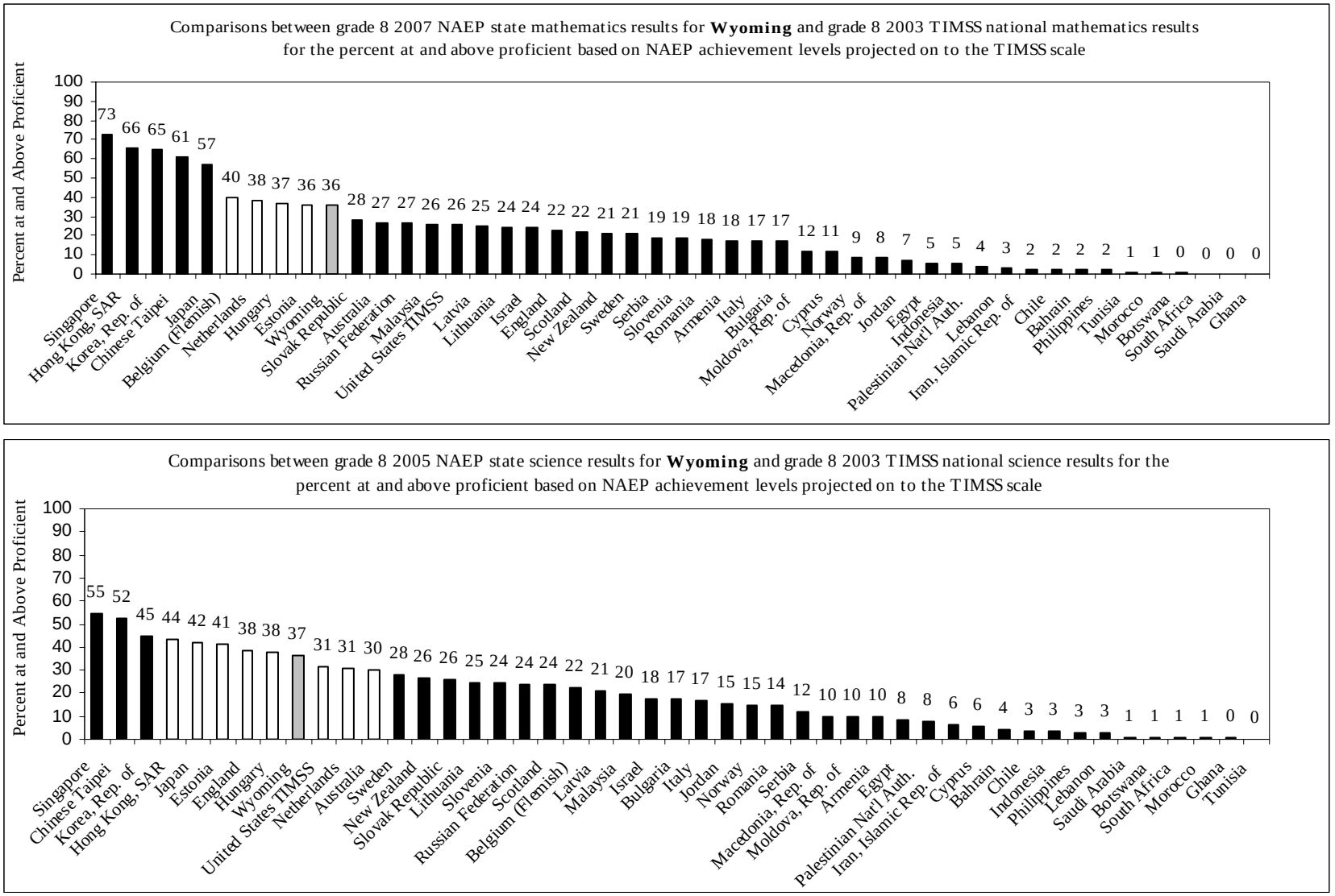
Figure 52: Wisconsin





Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

Figure 53: Wyoming



Source: Phillips, Gary W., *Chance Favors the Prepared Mind: Mathematics and Science Indicators For Comparing States and Nations*, AIR: Wash., DC, 2007.

Table 1: Achievement level of national mean on 2003 TIMSS grade 8 math scale (Basic-469, Proficient-566, Advanced-637)

Nation	Mean	Achievement Level of National Mean
Singapore	605	Proficient
Korea, Rep. of	589	Proficient
Hong Kong, SAR	586	Proficient
Chinese Taipei	585	Proficient
Japan	570	Proficient
Belgium (Flemish)	537	Basic
Netherlands	536	Basic
Estonia	531	Basic
Hungary	529	Basic
Latvia	508	Basic
Malaysia	508	Basic
Russian Federation	508	Basic
Slovak Republic	508	Basic
Australia	505	Basic
United States TIMSS	**504**	**Basic**
Lithuania	502	Basic
Sweden	499	Basic
England	498	Basic
Scotland	498	Basic
Israel	496	Basic
New Zealand	494	Basic
Slovenia	493	Basic
Italy	484	Basic
Armenia	478	Basic
Serbia	477	Basic
Bulgaria	476	Basic
Romania	475	Basic
Norway	461	Below Basic
Moldova, Rep. of	460	Below Basic
Cyprus	459	Below Basic
Macedonia, Rep. of	435	Below Basic
Lebanon	433	Below Basic
Jordan	424	Below Basic
Indonesia	411	Below Basic

Table 1: Achievement level of national mean on 2003 TIMSS grade 8 math scale (Basic-469, Proficient-566, Advanced-637)

Iran, Islamic Rep. of	411	Below Basic
Tunisia	410	Below Basic
Egypt	406	Below Basic
Bahrain	401	Below Basic
Palestinian Nat'l Auth.	390	Below Basic
Chile	387	Below Basic
Morocco	387	Below Basic
Philippines	378	Below Basic
Botswana	366	Below Basic
Saudi Arabia	332	Below Basic
Ghana	276	Below Basic
South Africa	264	Below Basic

Table 2: Achievement level of state/national mean on 2007 NAEP grade 8 math (Basic-262, Proficient-299, Advanced-333)

State/Nation	Mean	Achievement Level of National Mean
Massachusetts	298	Basic
Minnesota	292	Basic
North Dakota	292	Basic
Vermont	291	Basic
Kansas	290	Basic
New Jersey	289	Basic
South Dakota	288	Basic
Virginia	288	Basic
New Hampshire	288	Basic
Montana	287	Basic
Wyoming	287	Basic
Maine	286	Basic
Colorado	286	Basic
Pennsylvania	286	Basic
Texas	286	Basic
Maryland	286	Basic
Wisconsin	286	Basic
Iowa	285	Basic
DoDEA	285	Basic
Indiana	285	Basic
Washington	285	Basic
Ohio	285	Basic
North Carolina	284	Basic
Oregon	284	Basic
Nebraska	284	Basic
Idaho	284	Basic
Delaware	283	Basic
Alaska	283	Basic
Connecticut	282	Basic
South Carolina	282	Basic
Utah	281	Basic
Missouri	281	Basic
Illinois	280	Basic

Table 2: Achievement level of state/national mean on 2007 NAEP grade 8 math (Basic-262, Proficient-299, Advanced-333)

United States NAEP	**280**	**Basic**
New York	280	Basic
Kentucky	279	Basic
Florida	277	Basic
Michigan	277	Basic
Arizona	276	Basic
Rhode Island	275	Basic
Georgia	275	Basic
Oklahoma	275	Basic
Tennessee	274	Basic
Arkansas	274	Basic
Louisiana	272	Basic
Nevada	271	Basic
California	270	Basic
West Virginia	270	Basic
Hawaii	269	Basic
New Mexico	268	Basic
Alabama	266	Basic
Mississippi	265	Basic
District of Columbia	248	Below Basic

Table 3: Achievement level of national mean on 2003 TIMSS grade 8 science scale (Basic-494, Proficient-567, Advanced-670)

Nation	Mean	Achievement Level of National Mean
Singapore	578	Proficient
Chinese Taipei	571	Proficient
Korea, Rep. of	558	Basic
Hong Kong, SAR	556	Basic
Japan	552	Basic
Estonia	552	Basic
England	544	Basic
Hungary	543	Basic
Netherlands	536	Basic
United States TIMSS	**527**	**Basic**
Australia	527	Basic
Sweden	524	Basic
New Zealand	520	Basic
Slovenia	520	Basic
Lithuania	519	Basic
Slovak Republic	517	Basic
Belgium (Flemish)	516	Basic
Russian Federation	514	Basic
Scotland	512	Basic
Latvia	512	Basic
Malaysia	510	Basic
Norway	494	Basic
Italy	491	Below Basic
Israel	488	Below Basic
Bulgaria	479	Below Basic
Jordan	475	Below Basic
Moldova, Rep. of	472	Below Basic
Romania	470	Below Basic
Serbia	468	Below Basic
Armenia	461	Below Basic
Iran, Islamic Rep. of	453	Below Basic
Macedonia, Rep. of	449	Below Basic
Cyprus	441	Below Basic
Bahrain	438	Below Basic

Table 3: Achievement level of national mean on 2003 TIMSS grade 8 science scale (Basic-494, Proficient-567, Advanced-670)

Palestinian Nat'l Auth.	435	Below Basic
Egypt	421	Below Basic
Indonesia	420	Below Basic
Chile	413	Below Basic
Tunisia	404	Below Basic
Saudi Arabia	398	Below Basic
Morocco	396	Below Basic
Lebanon	393	Below Basic
Philippines	377	Below Basic
Botswana	365	Below Basic
Ghana	255	Below Basic
South Africa	244	Below Basic

Table 4: Achievement level of state/national mean on 2005 NAEP grade 8 science (Basic-143, Proficient-170, Advanced-208)

State/Nation	Mean	Achievement Level of National Mean
North Dakota	163	Basic
Montana	162	Basic
Vermont	162	Basic
New Hampshire	162	Basic
South Dakota	161	Basic
Massachusetts	161	Basic
DoDEA	160	Basic
Wyoming	159	Basic
Minnesota	158	Basic
Wisconsin	158	Basic
Idaho	158	Basic
Maine	158	Basic
Virginia	155	Basic
Ohio	155	Basic
Colorado	155	Basic
Michigan	155	Basic
Washington	154	Basic
Missouri	154	Basic
Utah	154	Basic
Oregon	153	Basic
New Jersey	153	Basic
Kentucky	153	Basic
Connecticut	152	Basic
Delaware	152	Basic
Indiana	150	Basic
Illinois	148	Basic
United States NAEP	**147**	**Basic**
West Virginia	147	Basic
Oklahoma	147	Basic
Rhode Island	146	Basic
South Carolina	145	Basic
Tennessee	145	Basic
Maryland	145	Basic
Arkansas	144	Basic

Table 4: Achievement level of state/national mean on 2005 NAEP grade 8 science (Basic-143, Proficient-170, Advanced-208)

North Carolina	144	Basic
Georgia	144	Basic
Texas	143	Basic
Florida	141	Below Basic
Arizona	140	Below Basic
Louisiana	138	Below Basic
Nevada	138	Below Basic
New Mexico	138	Below Basic
Alabama	138	Below Basic
Hawaii	136	Below Basic
California	136	Below Basic
Mississippi	132	Below Basic

References

Alexander-James Study Group, (1987), The nation's report card: Improving the assessment of student achievement. Cambridge, MA: National Academy of Education.

Beaton, A. E., and E. J. Gonzales. Comparing the NAEP Trial State Assessment Results with the IAEP International Results (report prepared for the National Academy of Education Panel on the NAEP Trial State Assessment). Stanford, CA: National Academy of Education, 1993.

Braswell, J. S., Lutkus, A. D., Grigg, W. S., Santapau, S. L., Tay-Lim, B. S. and Johnson, M. M., The Nation's Report Card: Mathematics 2000. Washington, DC: National Center for Education Statistics, 2001.

Cannell, J. J., "Nationally Normed Elementary Achievement Testing in America's Public Schools: How All 50 States Are above the National Average", Friends for Education, Daniels, WV, 1987.

Cannell, J. J. (1988), "Nationally Normed Elementary Achievement Testing in America's Public Schools: How All 50 States Are Above the National Average", Educational Measurement: Issues and Practice 7 (2), 5–9.

College Board, (1986), Press statement for release of 1986 SAT scores. New York: The College Board.

CRS Report RL 33434, (2006), Science, Technology, Engineering and Mathematics (STEM) Education Issues and legislative Options, by Kuenzi, J. J., Matthews, C. M. and Mangan, B. F.

Feinberg, L. (1988) "Above Average Has Become Testing Norm; States' Scores Called Misleadingly High". The Washington Post, Feb 8.

Fisher, G. M., (1992), "*The Development and History of the Poverty Thresholds*," Social Security Bulletin 55, no. 4 (Winter 1992):3-14

Fiske, E. (1988) "Standardized Test Scores: Voodoo Statistics", The New York Times, February 17.

GAO, Science, Technology, Engineering and Mathematics Trends and the Role of Federal programs: A Report to the Committee in Education and Workforce, House of Representatives, GAO-06-702T, Washington, D.C.: May 3, 2006.

Ginsburg, A. L., Noell, J. and Plisko, V. W., (1988), "Lessons from the Wall Chart,"

Educational Evaluation and Policy Analysis, 10, no. 1 (Spring, 1988), p. 1.).

Hauser, R.M, Edley, C.F. Jr., Koenig, J.A., and Elliott, S.W. (Eds.). Measuring Literacy: Performance Levels for Adults, Interim Report.Washington, DC: National Academies Press, 2005.

Johnson, E. G., and A. Siengondorf. Linking the National Assessment of Educational Progress and the Third International Mathematics and Science Study: Eighth Grade Results. (Publication No. NCES 98-500). Washington, DC: National Center for Education Statistics, 1998.

Johnson, E. G., J. Cohen, W.-H. Chen, T. Jiang, and Y. Zhang. 2000 NAEP–1999 TIMSS Linking Report. (Publication No. 2005-01). U.S. Department of Education. Washington DC: National Center for Education Statistics, 2005.

Kutner, M., Greenberg, E, and Baer, J. National Assessment of Adult literacy (NAAL): A First Look at the literacy of America's Adults in the 21st Century. National Center for Education Statistics, Institute of Education Sciences (ISE), U.S. Department of Education, NCES 2006-470, 2007.

LaPointe, A. E., N. A. Mead, and G. W. Phillips. A World of Differences: An International Assessment of Mathematics and Science. (Report No. 19-CAEP-01). Princeton, NJ: Educational Testing Service, 1989.

LaPointe, A. E., N. A. Mead, and J. M. Askew. Learning Mathematics. Princeton, NJ: Educational Testing Service, 1992.

Levine, D. B (editor), *Creating a Center for Education Statistics: A Time for Action*, National Academy of Sciences: National Academy Press, Washington, DC, 2006.

Linn, R. L. "Linking Results of District Assessments." Applied Measurement in Education, 6 (1993): 83–102.

Martin, M. O., Ina V. S. Mullis, E. J. Gonzalez, K. D. Gregory, T. A. Smith, S. J. Chrostowski, R. A. Garden, and K. M. O'Connor. TIMSS 1999 International Science Report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade. Chestnut Hill, MA: TIMSS and PIRLS International Study Center, Boston College. 2000.

Martin, M. O., Ina V. S. Mullis, E. J. Gonzalez, and S. J. Chrostowski. *TIMSS 2003 International Science Report: Findings From IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*, Chestnut Hill, MA: TIMSS and PIRLS International Study Center, Boston College, 2004.

Mislevy, R. J. Linking Educational Assessments: Concepts, Issues, Methods and Prospects. Princeton, NJ: Policy Information Center, Educational Testing Service, 1992.

Mullis, I. V. S., Dossey, J. A., Owen, E. H., and Phillips, G. W. (1991). The state of mathematics achievement: NAEP's I990 assessment of the nation and the Trial State Assessment. Washington, DC: National Center for Education Statistics.

Mullis, Ina V.S., M. O. Martin, E. J. Gonzales, K. D. Gregory, R. A. Garden, K. M. O'Connor, S. J. Chrostowski, and T. A. Smith. TIMSS 1999 International Mathematics Report: Findings from IEA's Repeat of The Third International Mathematics and Science Study at the Eighth Grade, Chestnut Hill, MA: TIMSS and PIRLS International Study Center, Boston College, 2000.

Mullis, Ina V. S., M. O. Martin, E. J. Gonzalez, and S. J. Chrostowski. *TIMSS 2003 International Mathematics Report: Findings From IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*, Chestnut Hill, MA: TIMSS and PIRLS International Study Center, Boston College, 2005.

National Center for Education Statistics. *Mapping 2005 State Proficiency Standards Onto the NAEP Scales* (NCES 2007–482). U.S. Department of Education, National Center for Education Statistics, Washington, D.C.: U.S. Government Printing Office

Nohara, D. A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R),and the Programme for International Student Assessment (PISA). (NCES Publication No. 2001-07).Washington, DC: U.S. Department of Education, National Center for Education Statistics, 2001.

O'Sullivan, C. Y., Lauko, M. A., Grigg, W. S. and Zhang, J., The Nation's Report Card: Science 2000. Washington, DC: National Center for Education Statistics, 2003.

National Science Foundation, Science and Engineering Indicators, 2006, Volume 1, Arlington, VA, NSB 06-01, January 2006.

Pashley, P. J. and Phillips, G. W. (1993), Toward World Class Standards: A Research Study Linking National and International Assessments. Princeton, NJ: Educational Testing Service.

Petersen, P. E. and Hess, Frederick, M. H. (2005), "Johnny Can Read…In Some States: Assessing the Rigor of State Assessment Systems", Education Next, No. 3.

Phillips, G. W., Finn, C. E. Jr., (1988), "The Lake Wobegon Effect: A Skeleton in the Testing Closet?", Educational Measurement: Issues and Practice, 7 (2), 10–11.

Phillips, G. W., (1990) "The Lake Wobegon Effect", Educational Measurement: Issues and Practice, 9 (3), pp 3&14.

Phillips, Gary W., (2007) Expressing International Educational Achievement in Terms of U.S. Performance Standards: Linking NAEP Achievement Levels to TIMSS, American Institutes for Research: Washington, DC.

Shavelson, R. J., McDonnell, L. & Oakes, J., (1991). What are educational indicators and indicator systems?. Practical Assessment, Research & Evaluation, 2(11).

U.S. Department of Education, National Center for Education Statistics, Digest of Education Statistics, 2004, NCES 2005-025, Oct. 2005, Table 169.

Vinovskis, M. A., (1999), The Road to Charlottesville: The 1989 Education Summit, National Education Goals Panel, Washington, DC.

Vu, P., (2007), "Lake Wobegon, U.S.A.--Where all the Children are Above Average: State of the States Report", STATELINE.ORG, January 31.

Wolter, K. Introduction to Variance Estimation. New York: Springer-Verlag, 1985.

Technical Appendix A: Statistical Linking NAEP to TIMSS

Linking

This appendix describes how and why the statistical linking between NAEP and TIMSS was done. Most of this appendix is reproduced from Phillips (2007).

Educators, researchers, and policymakers have considerable interest in how the American educational system compares to those in other countries. One major index for comparison is student academic achievement. Unfortunately, a lack of common metrics, as well as different definitions of performance standards, makes it difficult to compare measures of student achievement. The difficulty is similar to trying to compare the U.S. poverty level to that of other countries in the world. To do this, we first need a common metric. For example, we need to convert currencies of different countries to a common currency, such as dollars. Then we need a common definition and standard of poverty. That means either using a U.S. definition and standard and applying them to the rest of the world or using a common world definition and standard and applying those to the United States. No matter what common metric, definition, and standard are used, some people will argue it should have been done differently or not at all. Such comparisons are not perfect, always require more research, and should be done with caution. However, such cross-country comparisons result in the cross-fertilization of information and help inform debate. In general, comparisons are useful in providing information to policymakers and the general public to help them achieve broad understandings that they otherwise would not have.

This appendix shows how to link the scale of the *National Assessment of Educational Progress* (NAEP) to the scale of the *Third International Mathematics and Science Study* (TIMSS).[8] The purpose of this linking is to project the NAEP achievement levels onto the TIMSS scale. More specifically, the grade 8 NAEP: 2000 achievement levels in mathematics and science are projected on to the grade 8 TIMSS: 1999 assessment in mathematics and science. The linking equation is also applied to the 2003 TIMSS in mathematics and science. The goal is to project the grade 8 mathematics and science achievement levels in NAEP onto the TIMSS scale and thereby estimate the percent of basic, proficient, and advanced students in each country that participated in the 1999 TIMSS and 2003 TIMSS studies. The three achievement levels used were *basic*, *proficient*, and *advanced*, for both mathematics and science, as defined in *The Nation's Report Card: Mathematics 2000* (Braswell et al. 2001), and *The Nation's Report Card: Science 2000* (O'Sullivan et al. 2003), respectively. The TIMSS results may be found in *TIMSS 1999: International Mathematics Report* (Mullis et al. 2000), *TIMSS 1999: International Science Report* (Martin et al. 2000), *TIMSS 2003: International Mathematics Report* (Mullis et al. 2005), and *TIMSS 2003: International Science Report* (Martin et al. 2004).

Linking Methods

Mislevy (1992) and Linn (1993) have described many of the conceptual and statistical issues associated with linking assessments. They have outlined four forms of statistical linking: equating, calibration, projection, and statistical moderation. A further explication of the differences is provided here.

[8] The definition of the acronym TIMSS was subsequently changed to Trends in International Mathematics and Science Study.

The three assumptions that distinguish the different forms of statistical linking are that two tests (call them *X* and *Y*) have true scores that are highly correlated, measure the same content, and are equally reliable. These assumptions are displayed in Table 5

Table 5: Statistically linking test X and test Y

	Equating	Calibration	Projection	Moderation
High true score correlation	x[9]	X[8]	x	
Same content	x	x		
Equal reliability	x			

In *equating,* both tests, *X* and *Y,* have been designed and developed to be equally reliable, and each measures the same content. Equating is used when the goal is to relate two alternate forms of the same test, such as alternate forms of the ACT or the SAT. Under these conditions, the only difference between the two tests is the metric, such as expressing temperature in terms of Fahrenheit or Celsius. In equating the distributions of test *X* and *Y* are aligned or matched up directly. The matching can be done with equipercentile equating or linear equating, and the distributions can be either observed score distributions or estimates of the true score distributions. When the three assumptions (high correlation, same content, and equal reliability) are met:

- the linking function should be the same for *X* expressed in terms of *Y,* and for *Y* expressed in terms of *X*, and
- the linking function should be the same for different subgroups, across contexts and time.

In *calibration* (for example with the use of item-response theory), two tests are assumed to measure the same content, but they are not equally reliable. For example, one test *X* might be a long test whereas the other test *Y* is short. The two versions of the test are not equated, but they are indirectly comparable because they have been calibrated to a common scale θ. This type of linking is done across grades and across years in NAEP, TIMSS, most state criterion-referenced tests, and most nationally standardized norm-referenced tests. Calibration procedures provide unbiased estimates for individual students and means, but additional statistical machinery is needed to accurately estimate group characteristics such as the variance or the percent at and above achievement levels. When the two assumptions (high correlation and same content) are met:

- the linking function between *X* and θ (e.g., the test characteristic curve) is different from the linking function between *Y* and θ,
- both *X* and *Y* can be used to get unbiased estimates of θ for individual students (although the error in the estimates will be higher for *Y*), however
- the observed score distributions of *X* for groups do not match the observed score distributions for *Y*.

In *projection,* a regression equation uses the correlation between the two tests to predict the scores on one test *Y* from those of another test *X*. There is no assumption that the two tests

[9] The true-score correlation between *X* and *Y* is assumed to equal 1.0.

measure the same content or that they are equally reliable. With projection, there is no longer a symmetric relationship between one test and the other. The conversion table for predicting the first test from the second is different from the table predicting the second test from the first. When the assumption of high correlation is met:

- the linking function for *X* expressed in terms of *Y* (e.g., regression equation) will be different from the linking function for *Y* expressed in terms of *X*, and
- the linking function will likely be different for different subgroups, across contexts and time.

In *statistical moderation*, the scores on the first test *X* are adjusted to have the same distributional characteristics as the scores on the second test *Y*. In this case *X* is linked to *Y*. This is typically done by matching the means and standard deviations of *X* and *Y*, or matching their percentile ranks. The usual assumption is that both, *X* and *Y*, have been administered to comparable populations of students (e.g., the student populations taking both tests are randomly equivalent). Statistical moderation typically does not use the correlation between the two tests. When statistical moderation is used:

- the linking function for *X* expressed in terms of *Y* (e.g., a z-score equivalency) will be different from the linking function for *Y* expressed in terms of *X*,
- the linking function will likely be different for different subgroups, across contexts and time, and
- the degree of the relationship between *X* and *Y* is typically unknown.

Linking is essentially a process that provides a concordance table that expresses scores on one test (e.g., TIMSS) in terms of the metric of another test (e.g., NAEP). This paper uses *statistical moderation* to link the NAEP achievement levels to TIMSS by extending the process used in the *2000 NAEP–1999 TIMSS Linking Report* (Johnson et al. 2005). This extension was an extremely easy process because that report did all the hard work. The main goal of the report (Johnson et al. 2005) was to use the link between NAEP and TIMSS to estimate how the students in the states of the United States would have performed if they had taken the TIMSS test, based on the fact they took the NAEP test. This same linking process also can be used to answer the question, "How would other countries perform if their TIMSS results could be expressed in terms of NAEP achievement levels?" In other words, we can use the findings in the 2005 report by Johnson and colleagues to project the NAEP achievement levels onto the TIMSS scale as a way to interpret how each country performed on the TIMSS assessment in terms of U.S. performance standards. This paper takes that approach.

Linking NAEP to International Assessments

Several major attempts have been made to link NAEP statistically to international assessments.

The *first* attempt involved linking the 1991 International Assessment of Educational Progress (IAEP) to the 1992 NAEP in mathematics (Pashley and Phillips, 1993). The IAEP was first conducted in February 1988 in five countries (Ireland, Korea, Spain, the United Kingdom, and the United States) and four provinces in Canada (LaPointe, Mead, and Phillips, 1989) using representative samples of 13-year old students assessed in mathematics and science. The IAEP was expanded and repeated again in 1991 (LaPointe, Meade, and Askew, 1992) in 20 countries in which representative samples of 9- and 13-year old students were assessed in mathematics and

science. Pashley and Phillips (1993) conducted the IAEP-NAEP linking study in mathematics using *projection* methodology. In order to establish the link between the IAEP and NAEP, a nationally representative linking sample of 1,609 students was administered both the IAEP and NAEP in 1992. The linking study used samples of 8th-grade students who took NAEP versus 13-year-old students who took the IAEP (NAEP was based on grade whereas the IAEP was based on age). The direction of the link was to predict NAEP performance from IAEP results in other countries. The purpose of the study was to estimate how other countries stacked up against the NAEP achievement levels. The IAEP-NAEP linkage was done within the context of the policy environment at the time. The nation's governors, along with the President had held the National Education Summit and adopted six broad national goals. The fourth goal was that, by the year 2000, "U.S. students would be the first in the world in science and mathematics achievement." The IAEP-NAEP linking study was the first effort to address directly the need for a common metric and common standard in international comparisons (i.e., predict how other countries would do on NAEP based on their performance on IAEP). Once the predicted NAEP scores were obtained, then the NAEP achievement levels were used to report different countries' performance. The IAEP was not repeated; however, it had many design features (such as linking studies) that were incorporated into subsequent international assessments of TIMSS.

A *second* attempt to link NAEP to an international study was done by Beaton and Gonzales (1993). They used *statistical moderation* to link the 1991 IAEP to the 1990 NAEP scale in mathematics. The results of the Beaton and Gonzales (1993) study were similar to the Pashley and Phillips (1993) study only for countries with performance similar to the U.S. average.

The *third* study used *statistical moderation* to link the grade 4 and grade 8 1996 NAEP to 1995 TIMSS, grades 4 and 8, mathematics and science (Johnson and Siengondorf, 1998). Based on the validation analyses (in two states that took both NAEP and TIMSS), the NAEP-TIMSS link appeared to work at grade 8 but not at grade 4.[10]

The *fourth* study (Johnson et al. 2005) used *projection* methods (similar to Pashley and Phillips, 1993) for grade 8 mathematics and science to link NAEP to TIMSS. The TIMSS assessment in mathematics and science was conducted in 1999, and the NAEP assessment in math and science was conducted in 2000. In addition to projection methods, the study also used *statistical moderation* as a secondary method of linking. Based on a validation study in which 12 states took both NAEP and TIMSS, the general finding was that, for the U.S. national linking sample, the projection method did not work. However, the statistical moderation method (which used the national samples of both NAEP and TIMSS instead of the linking sample) did perform well in the validation study.

Although statistical moderation provided an acceptable link, this approach is considered the weakest linking method because it does not use the correlation between the two assessments. In this case, however, it is the only method available so far that appears to work for linking NAEP to TIMSS. The estimates provided by statistical moderation should be considered rough, ballpark estimates and should be used only for broad policy understandings.

[10] The link worked at grade 8 based on the validation sample. The predicted TIMSS results for Minnesota (the only state that administered the 8th grade TIMSS) were comparable to the actual TIMSS results. The link did not work at grade 4. The predicted TIMSS results for the two states that administered 4th-grade TIMSS (Colorado and Minnesota) were considerably higher than the actual TIMSS results. The study was not able to determine why this result occurred in the grade 4 link.

Linking NAEP Achievement Levels to TIMSS

This report used statistical moderation for randomly equivalent populations. The main purpose of the NAEP-TIMSS link by Johnson and colleagues (2005) was to predict TIMSS results for the states within the United States, based on their performance on NAEP. The current paper re-analyses the data provided by that study to extend this process and link NAEP achievement levels to TIMSS. This analysis provides estimates of how countries outside the United States that participated in the TIMSS would perform, using the NAEP achievement levels estimated on the TIMSS scale.

Caveats

Several important caveats are associated with these analyses. *First*, the standard errors and the validation analyses are based on data collected only within the United States. In the United States, students took both NAEP and TIMSS; in all other countries, however, students only took TIMSS. Whether the linking parameters are stable in other countries is an empirical question that the study by Johnson and colleagues (2005) could not answer. In fact, no international linking study has been designed to answer this question. There is no guarantee that linking parameters estimated from one group (e.g., the United States) will be the same in other groups.

The *second* caveat is that the percentage at or above basic, proficient, and advanced levels in the tables below is based on the assumption of a "normal distribution" of performance within each country. In most cases, this assumption should be approximately true.

The *third* caveat is that this paper used the linking parameters obtained from the 2000 NAEP and 1999 TIMSS to estimate achievement levels in the subsequent 2003 TIMSS; that is, the linking parameters are assumed to be stable across years. More than likely, they are not stable across years; nevertheless, they should be sufficient for very rough approximations. A better approach would be using a linking study that explicitly used the 2003 TIMSS. Because no linking study was conducted during the administration of the 2003 TIMSS, the past 1999–2000 study is all that is available. In fact, no linking studies have been conducted after the 2000 NAEP and 1999 TIMSS assessments.

The *fourth* caveat is that the achievement levels developed for the NAEP were based on the content of the NAEP. Although content similarities between the 8th-grade NAEP and TIMSS (Nohara, 2001) are substantial, the NAEP achievement levels do not strictly apply to TIMSS. The problem is similar to the poverty-level analogy used above. Definitions and standards of poverty in the United States will not strictly apply to other countries in the world; however, the definitions and standards can be used to estimate approximately how the rest of the world relates to U.S. expectations of a decent standard of living. For a thoughtful and thorough discussion of similarities and differences in several international assessments the reader should review the report at http://nces.ed.gov/Surveys/PISA/pdf/comppaper12082004.pdf.

All of these caveats reinforce what was said above about the limits of inference from these data. At best, these concordance tables should be used for rough approximations to give policy makers a general idea of how the United States stacks up with the rest of the world.

Linking Using Statistical Moderation

Basic Equations

In the study by Johnson and colleagues (2005), NAEP was linked to TIMSS by using statistical moderation. This means the estimated $\hat{T}IMSS$ scores are actually NAEP scores adjusted to have the same mean and standard deviation as TIMSS. That is what it means in *statistical moderation* to say "NAEP is linked to TIMSS." In the present study the same data were re-analyzed to link the NAEP achievement levels to the TIMSS scale. The estimated $\hat{T}IMSS$ score associated with a NAEP achievement level ($\hat{T}IMSS_{level}$) is

$$\hat{T}IMSS_{level} = \hat{A} + \hat{B}\left(NAEP_{level}\right). \qquad (0.1)$$

In equation (0.1) $\hat{A}$ is an estimate of the intercept of a straight line, and $\hat{B}$ is an estimate of the slope defined by

$$\hat{A} = \hat{\mu}_{TIMSS} - \hat{B}\hat{\mu}_{NAEP}$$
$$\hat{B} = \frac{\hat{\sigma}_{TIMSS}}{\hat{\sigma}_{NAEP}}. \qquad (0.2)$$

In equation (0.2), $\hat{\mu}_{NAEP}$ and $\hat{\mu}_{TIMSS}$ are the national means of the U.S. NAEP and TIMSS results for public school students, respectively, while $\hat{\sigma}_{NAEP}$ and $\hat{\sigma}_{TIMSS}$ are the standard deviations of the tests. The means and standard deviations in equation (0.2) are reported in table 6. The resulting estimates of the linking parameters $\hat{A}$ and $\hat{B}$ are reported in table 7.

Table 6: Means and standard deviations for national samples of grade 8 U.S. public school students, 1999 TIMSS and 2000 NAEP

Subject	TIMSS		NAEP	
	Mean	SD	Mean	SD
Mathematics	498.2	88.4	274.4	37.4
Science	510.4	98.0	149.2	36.2

SOURCES: National data file from the *1999 IEA Trends in International Mathematics and Science Study* (TIMSS-99) and the *2000 National Assessment of Educational Progress* (NAEP).

Table 7: Estimating 1999 TIMSS scores from 2000 NAEP, using statistical moderation with U.S. national samples

Subject	A	B
Mathematics	–150.38	2.36
Science	106.49	2.71

The NAEP achievement levels projected on to the TIMSS scale are reported in table 8 for mathematics and table 9 for science. The details of the estimation procedure for the standard error of the projected achievement levels are presented in the next section of this technical appendix.

Table 8: Grade 8 2000 NAEP mathematics achievement levels linked to grade 8 1999-TIMSS mathematics

	NAEP Achievement Level	NAEP Achievement Level Projected on to the TIMSS Scale	Standard Error of Linking for Projected Achievement Level
Basic	262	469	4.83
Proficient	299	556	5.13
Advanced	333	637	6.72

Table 9: Grade 8 2000 NAEP science achievement levels linked to grade 8 1999-TIMSS science

	NAEP Achievement Level	NAEP Achievement Level Projected on to the TIMSS Scale	Standard Error of Linking for Projected Achievement Level
Basic	143	494	5.44
Proficient	170	567	5.59
Advanced	208	670	6.63

Linking Error Variance

The linking procedure described in this paper is straightforward and easy to accomplish. The intermediate calculations of the error variance, however, are complex and tedious. This appendix describes the details of how the error variances reported in the paper were determined. Most of these analyses, especially those involving plausible values, were done as part of the study by Johnson et al. (2005). Furthermore, the analyses of plausible values have been well documented in the various technical manuals of both NAEP and TIMSS.

With statistical moderation, the estimated $\hat{T}IMSS_{level}$ is a linear transformation of $NAEP_{level}$. Therefore, the error variance in $\hat{T}IMSS_{level}$ is

$$\hat{\sigma}^2_{\hat{T}IMSS_{level}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{level}} + \hat{\sigma}^2_A + 2\left(NAEP_{level}\right)\hat{\sigma}_{AB} + \left(NAEP_{level}\right)^2\hat{\sigma}^2_B. \qquad (0.3)$$

According to Johnson et al. (2005), the error variances of the parameters of the linear transformation, $\hat{\sigma}_A^2$, $2\hat{\sigma}_{AB}^2$ and $\hat{\sigma}_B^2$ can be approximated by Taylor-series linearization (Wolter, 1985).

$$\hat{\sigma}_A^2 = \hat{B}^2\hat{\sigma}_{\mu_{NAEP}}^2 + \hat{\sigma}_{\mu_{TIMSS}}^2 + \hat{\mu}_{NAEP}^2\hat{B}^2\left[\frac{Var\left(\hat{\sigma}_{TIMSS}\right)}{\hat{\sigma}_{TIMSS}^2} + \frac{Var\left(\hat{\sigma}_{NAEP}\right)}{\hat{\sigma}_{NAEP}^2}\right]$$

$$2\hat{\sigma}_{AB} = -2\hat{\mu}_{NAEP}\hat{B}^2\left[\frac{Var\left(\hat{\sigma}_{TIMSS}\right)}{\hat{\sigma}_{TIMSS}^2} + \frac{Var\left(\hat{\sigma}_{NAEP}\right)}{\hat{\sigma}_{NAEP}^2}\right] \qquad (0.4)$$

$$\hat{\sigma}_B^2 = \hat{B}^2\left[\frac{Var\left(\hat{\sigma}_{TIMSS}\right)}{\hat{\sigma}_{TIMSS}^2} + \frac{Var\left(\hat{\sigma}_{NAEP}\right)}{\hat{\sigma}_{NAEP}^2}\right].$$

In this particular application, we can treat the NAEP achievement levels as fixed, so there is no error associated with $NAEP_{level}$, therefore $\hat{B}^2\hat{\sigma}_{NAEP_{level}}^2 = 0$. Equations (0.3) and(0.4), along with the data provided by Johnson et al. (2005), were used to derive the estimates in this paper.[11] The estimated achievement levels (along with their linking errors) are presented in table 3 for TIMSS mathematics and table 4 for TIMSS science. The standard error of linking reported in table 3 and table 4 is the square root of equation (0.3). The intermediate calculations for equations (0.3) and (0.4) are presented below.

Parameter estimates of the mean and standard deviation

The process begins with the analysis of plausible values for both NAEP and TIMSS. In both NAEP and TIMSS, five plausible values are used to represent the student's posterior distribution. Let us label the parameter we are estimating as "*t,*" and the number of plausible values as "*M,*" and the estimates of *t* as $\hat{t}_m$, for $m = 1, 2, ..M$. The average of the statistics is t^*, where $t^* = \sum_{m=1}^{M}\frac{t_m}{M}$.

Tables 10A and 10B are the calculations for the parameter estimates of the means and standard deviations (SD).

Table 10A: Estimating the mean and standard deviation in U.S. national samples (public schools) for grade 8 mathematics

	Plausible value 1	Plausible value 2	Plausible value 3	Plausible value 4	Plausible value 5	Mean plausible value (*t**)
2000 NAEP mathematics mean	274.505	274.467	274.329	274.297	274.480	274.416
1999 TIMSS mathematics mean	498.505	498.378	497.883	497.742	498.671	498.236
2000 NAEP mathematics SD	37.482	37.305	37.337	37.217	37.433	37.355
1999 TIMSS mathematics SD	86.481	88.451	89.410	89.047	88.549	88.388

[11] I wish to thank Tao Jiang at the American Institutes for Research® for providing the plausible values results for both NAEP and TIMSS from the study (Johnson et al. 2005) that allowed for the calculation of standard errors in this paper.

Table 10B: Estimating the mean and standard deviation in U.S. national samples (public schools) for grade 8 science

	Plausible value 1	Plausible value 2	Plausible value 3	Plausible value 4	Plausible value 5	Mean plausible value (t^*)
2000 NAEP science mean	149.301	149.229	148.998	149.037	149.382	149.189
1999 TIMSS science mean	509.305	510.657	510.460	509.437	512.086	510.389
2000 NAEP science SD	36.212	36.354	36.020	36.173	36.354	36.222
1999 TIMSS science SD	97.490	98.647	96.803	98.276	98.643	97.972

Error variance (sampling) of the mean and standard deviation

The error variances for the parameter estimates in Tables 10A and 10B each have two components—error variance due to sampling (U^*) and error variance due to measurement (B^*). The sampling error in the estimates of the means and standard deviations were obtained by using a jackknife error variance approach for complex samples. The jackknife procedure was carried out for each plausible value and then averaged across all five plausible values. In the jackknife procedure, one primary sampling unit (PSU) is excluded; the sampling weights are redistributed across the other units within the stratum in which the PSU was excluded; the mean and standard deviation are calculated on the remaining PSUs; and the process is repeated until all PSUs have been excluded. After the jackknife procedure is carried out on each plausible value, the average across plausible values is $U^* = \sum_{m=1}^{M} \frac{U_m}{M}$.

This process resulted in the variance estimates reported in Tables 11A and 11B which are estimates of error variance due to sampling for the means and standard deviations.

Table 11A: Sampling error variance of the mean and standard deviation (U^*) for grade 8 mathematics

Variance of NAEP mean 2000 mathematics from jackknife	0.640
Variance of TIMSS mean 1999 mathematics from jackknife	18.490
Variance of NAEP SD 2000 mathematics from jackknife	0.250
Variance of TIMSS SD 1999 mathematics from jackknife	6.250

Table 11B: Sampling error variance of the mean and standard deviation (U^*) for grade 8 science

Variance of NAEP mean 2000 science from jackknife	0.490
Variance of TIMSS mean 1999 science from jackknife	25.000
Variance of NAEP SD 2000 science from jackknife	0.250
Variance of TIMSS SD 1999 science from jackknife	4.410

Error variance (measurement) of the mean and standard deviation

The error variance due to measurement is estimated by the variance between plausible values. This is estimated by $B^* = \frac{1+(1/M)}{M-1}\sum_{m=1}^{M}(t_m - t^*)^2$. The error variance due to measurement is in Tables 12A and 12B.

Table 12A: Measurement error variance of the mean and standard deviation (B^*) for grade 8 mathematics

Variance of NAEP mean 2000 mathematics from plausible values	0.011
Variance of TIMSS mean 1999 mathematics from plausible values	0.195
Variance of NAEP SD 2000 mathematics from plausible values	0.013
Variance of TIMSS SD 1999 mathematics from plausible values	1.544

Table 12B: Measurement error variance of the mean and standard deviation (B^*) for grade 8 science

Variance of NAEP mean 2000 science from plausible values	0.033
Variance of TIMSS mean 1999 science from plausible values	1.511
Variance of NAEP SD 2000 science from plausible values	0.023
Variance of TIMSS SD 1999 science from plausible values	0.779

Error variance (total) of the mean and standard deviation

The total error variance is $V^* = U^* + B^*$ and is contained in Tables 13A and 13B.

Table 13A: Total error variance of the mean and standard deviation (V^*) for grade 8 mathematics

Variance of NAEP mean 2000 mathematics	0.651
Variance of TIMSS mean 1999 mathematics	18.685
Variance of NAEP SD 2000 mathematics	0.263
Variance of TIMSS SD 1999 mathematics	7.794

Table 13B: Total error variance of the mean and standard deviation (V^*) for grade 8 science

Variance of NAEP mean 2000 science	0.523
Variance of TIMSS mean 1999 science	26.511
Variance of NAEP SD 2000 science	0.273
Variance of TIMSS SD 1999 science	5.189

Parameter estimates of the linking parameters *A* and *B*

The linking parameters are then calculated for each plausible value, using equation (0.2). The linking parameter estimates are then averaged over the five plausible values as reported in Tables 14A and 14B.

Table 14A: Estimating the linking parameters A and B in the U.S. national samples (public schools) for grade 8 mathematics

	Plausible value 1	Plausible value 2	Plausible value 3	Plausible value 4	Plausible value 5	Mean plausible value (t^*)
$\hat{A}$	–134.854	–152.393	–159.041	–158.554	–150.619	–151.077
$\hat{B}$	2.307	2.371	2.395	2.393	2.366	2.366

Table 14B: Estimating the linking parameters A and B in the U.S. national samples (public schools) for grade 8 science

	Plausible value 1	Plausible value 2	Plausible value 3	Plausible value 4	Plausible value 5	Mean plausible value (*t**)
$\hat{A}$	107.351	105.720	110.029	104.531	106.752	106.877
$\hat{B}$	2.692	2.714	2.688	2.717	2.713	2.705

Error variance (sampling) of the linking parameters *A* and *B*

The error variance of the linking parameters estimates $\hat{A}$ and $\hat{B}$ is found by equation (0.4). The linking error variance also has two components—one due to sampling and one due to measurement error. The quantities needed to estimate the error variance in the linking parameters due to sampling are contained in Tables 11A and 11B. The quantities needed to estimate the error variance in the linking parameters due to measurement error are contained in Tables 12A and 12B. Substituting the estimates in Tables 11A and 11B in equation (0.4), we have the error variance in the linking parameters due to sampling. These are reported in Tables 15A and 15B.

Table 15A: Sampling error variance in NAEP–TIMSS linking parameters for mathematics

Error variance in A, $\left(\hat{\sigma}^2_{A(s)}\right)$	434.901
Two times the covariance between A and B, $2\left(\hat{\sigma}_{AB(s)}\right)$	–3.009
Error variance in B, $\left(\hat{\sigma}_{B(s)}\right)$	0.005

Table 15B: Sampling error variance in NAEP–TIMSS linking parameters for science

Error variance in A, $\left(\hat{\sigma}^2_{A(s)}\right)$	108.740
Two times the covariance between A and B, $2\left(\hat{\sigma}_{AB(s)}\right)$	–1.086
Error variance in B, $\left(\hat{\sigma}_{B(s)}\right)$	0.004

Error variance (measurement) of the linking parameters *A* and *B*

Substituting the estimates in Tables 12A and 12B in equation (0.4) provides the error variance in the linking parameters due to measurement error, as reported in Tables 16A and 16B.

Table 16A: Measurement error variance in NAEP–TIMSS linking parameters for grade 8 mathematics

Error variance in A, $\left(\hat{\sigma}^2_{A(m)}\right)$	87.575
Two times the covariance between A and B, $2\left(\hat{\sigma}_{AB(m)}\right)$	–0.636
Error variance in B, $\left(\hat{\sigma}_{B}\right)$	0.001

Table 16B: Measurement error variance in NAEP–TIMSS linking parameters for grade 8 science

Error variance in A, $\left(\hat{\sigma}^2_{A(m)}\right)$	14.040
Two times the covariance between A & B, $2\left(\hat{\sigma}_{AB(m)}\right)$	–0.165
Error variance in B, $\left(\hat{\sigma}_{B(m)}\right)$	0.001

Error variance (total) of the linking parameters *A* and *B*

The sum of the sampling error variances in Tables 15A and 15B and the measurement error variances in Tables 16A and 16B yield the total error variances in the linking parameters reported in Tables 17A and 17B.

Table 17A: Total error variance in NAEP–TIMSS linking parameters for grade 8 mathematics

Error variance in A, $\left(\hat{\sigma}^2_{A}\right)$	522.476
Two times the covariance between A and B, $2\left(\hat{\sigma}_{AB}\right)$	–3.645
Error variance in B, $\left(\hat{\sigma}_{B}\right)$	0.007

Table 17B: Total error variance in NAEP–TIMSS linking parameters for grade 8 science

Error variance in A, $\left(\hat{\sigma}^2_{A}\right)$	122.781
Two times the covariance between A and B, $2\left(\hat{\sigma}_{AB}\right)$	–1.251
Error variance in B, $\left(\hat{\sigma}_{B}\right)$	0.004

Error variance (sampling) of the projected NAEP achievement levels

The linking error variance of the projected NAEP achievement levels on the TIMSS scale is found in equation (0.3). The linking error variance also has two components—one due to sampling, and one due to measurement error. The quantities needed to estimate the error variance in the projected achievement levels due to sampling are contained in Tables 15A and 15B. The quantities needed to estimate the error variance in the linking parameters due to measurement error are contained in Tables 16A and 16B. Substituting the estimates in Tables 15A and 15B in equation (0.3), we have the linking error variance in the projected achievement levels due to sampling. These are reported in Tables 18A and 18B.[12]

Table 18A: Error variance in linking due to sampling for NAEP achievement levels projected onto TIMSS grade 8 mathematics scale

$\hat{\sigma}^2_{TIMSS_{basic}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{basic}} + \hat{\sigma}^2_{A(s)} + 2\left(NAEP_{basic}\right)\hat{\sigma}_{AB(s)} + \left(NAEP_{basic}\right)^2\hat{\sigma}^2_{B(s)}$	22.918
$\hat{\sigma}^2_{TIMSS_{prof}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{prof}} + \hat{\sigma}^2_{A(s)} + 2\left(NAEP_{prof}\right)\hat{\sigma}_{AB(s)} + \left(NAEP_{prof}\right)^2\hat{\sigma}^2_{B(s)}$	25.387
$\hat{\sigma}_{TIMSS_{adv}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{adv}} + \hat{\sigma}^2_{A(s)} + 2\left(NAEP_{adv}\right)\hat{\sigma}_{AB(s)} + \left(NAEP_{adv}\right)^2\hat{\sigma}^2_{B(s)}$	40.889

Table 18B: Error variance in linking due to sampling for NAEP achievement levels projected onto TIMSS grade 8 science scale

$\hat{\sigma}^2_{TIMSS_{basic}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{basic}} + \hat{\sigma}^2_{A(s)} + 2\left(NAEP_{basic}\right)\hat{\sigma}_{AB(s)} + \left(NAEP_{basic}\right)^2\hat{\sigma}^2_{B(s)}$	27.883
$\hat{\sigma}^2_{TIMSS_{prof}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{prof}} + \hat{\sigma}^2_{A(s)} + 2\left(NAEP_{prof}\right)\hat{\sigma}_{AB(s)} + \left(NAEP_{prof}\right)^2\hat{\sigma}^2_{B(s)}$	29.319
$\hat{\sigma}_{TIMSS_{adv}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{adv}} + \hat{\sigma}^2_{A(s)} + 2\left(NAEP_{adv}\right)\hat{\sigma}_{AB(s)} + \left(NAEP_{adv}\right)^2\hat{\sigma}^2_{B(s)}$	40.330

Error variance (measurement) of the projected NAEP achievement levels

Substituting the estimates in Tables 16A and 16B in equation (0.3) provides the linking error variance in the projected achievement levels due to measurement error as reported in Tables 19A and 19B.

Table 19A: Error variance in linking due to measurement for NAEP achievement levels projected onto TIMSS grade 8 mathematics scale

$\hat{\sigma}^2_{TIMSS_{basic}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{basic}} + \hat{\sigma}^2_{A(m)} + 2\left(NAEP_{basic}\right)\hat{\sigma}_{AB(m)} + \left(NAEP_{basic}\right)^2\hat{\sigma}^2_{B(m)}$	0.435
$\hat{\sigma}^2_{TIMSS_{prof}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{prof}} + \hat{\sigma}^2_{A(m)} + 2\left(NAEP_{prof}\right)\hat{\sigma}_{AB(m)} + \left(NAEP_{prof}\right)^2\hat{\sigma}^2_{B(m)}$	0.957
$\hat{\sigma}_{TIMSS_{adv}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{adv}} + \hat{\sigma}^2_{A(m)} + 2\left(NAEP_{adv}\right)\hat{\sigma}_{AB(m)} + \left(NAEP_{adv}\right)^2\hat{\sigma}^2_{B(m)}$	4.236

[12] Since the NAEP achievement levels are a known parameter, we assume throughout this paper that $\hat{B}^2\hat{\sigma}^2_{NAEP_{ach\ level}}$ is equal to zero.

Table 19B: Error variance in linking due to measurement for NAEP achievement levels projected onto TIMSS grade 8 science scale

$\hat{\sigma}^2_{TIMSS_{basic}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{basic}} + \hat{\sigma}^2_{A(m)} + 2\left(NAEP_{basic}\right)\hat{\sigma}_{AB(m)} + \left(NAEP_{basic}\right)^2\hat{\sigma}^2_{B(m)}$	1.719
$\hat{\sigma}^2_{TIMSS_{prof}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{prof}} + \hat{\sigma}^2_{A(m)} + 2\left(NAEP_{prof}\right)\hat{\sigma}_{AB(m)} + \left(NAEP_{prof}\right)^2\hat{\sigma}^2_{B(m)}$	1.938
$\hat{\sigma}_{TIMSS_{adv}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{adv}} + \hat{\sigma}^2_{A(m)} + 2\left(NAEP_{adv}\right)\hat{\sigma}_{AB(m)} + \left(NAEP_{adv}\right)^2\hat{\sigma}^2_{B(m)}$	3.616

Error variance (total) of the projected NAEP achievement levels

The sum of the linking error variance due to sampling in Tables 18A and 18B and the linking error variance due to measurement Tables 19A and 19B yields the total linking error variances in the projected achievement levels on the TIMSS scale reported in Tables 20A and 20B.

Table 20A: Total error variance in linking for NAEP achievement levels projected onto TIMSS grade 8 mathematics scale

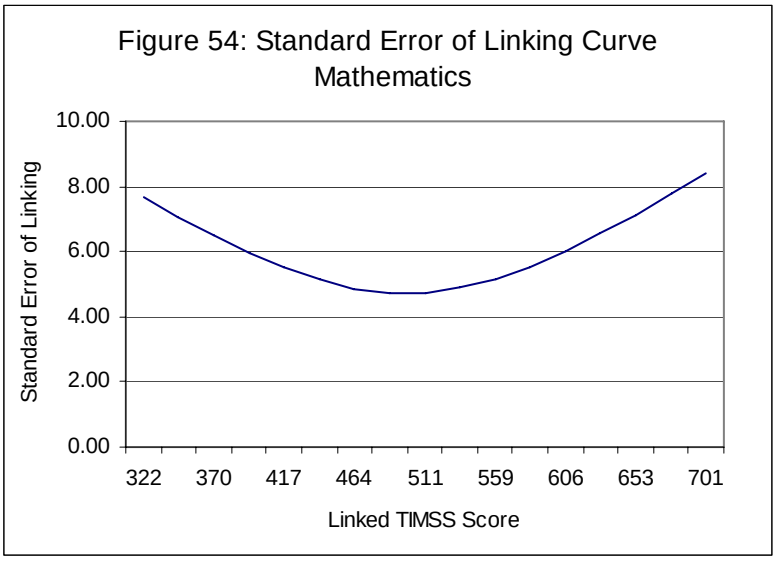
$\hat{\sigma}^2_{TIMSS_{basic}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{basic}} + \hat{\sigma}^2_{A} + 2\left(NAEP_{basic}\right)\hat{\sigma}_{AB} + \left(NAEP_{basic}\right)^2\hat{\sigma}^2_{B}$	23.353
$\hat{\sigma}^2_{TIMSS_{prof}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{prof}} + \hat{\sigma}^2_{A} + 2\left(NAEP_{prof}\right)\hat{\sigma}_{AB} + \left(NAEP_{prof}\right)^2\hat{\sigma}^2_{B}$	26.343
$\hat{\sigma}_{TIMSS_{adv}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{adv}} + \hat{\sigma}^2_{A} + 2\left(NAEP_{adv}\right)\hat{\sigma}_{AB} + \left(NAEP_{adv}\right)^2\hat{\sigma}^2_{B}$	45.124

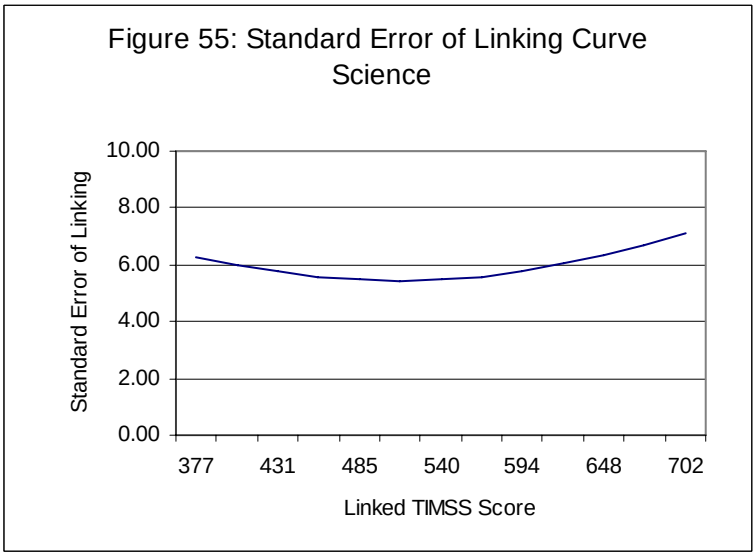
Table 20B: Total error variance in linking for NAEP achievement levels projected onto TIMSS grade 8 science scale

$\hat{\sigma}^2_{TIMSS_{basic}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{basic}} + \hat{\sigma}^2_{A} + 2\left(NAEP_{basic}\right)\hat{\sigma}_{AB} + \left(NAEP_{basic}\right)^2\hat{\sigma}^2_{B}$	29.602
$\hat{\sigma}^2_{TIMSS_{prof}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{prof}} + \hat{\sigma}^2_{A} + 2\left(NAEP_{prof}\right)\hat{\sigma}_{AB} + \left(NAEP_{prof}\right)^2\hat{\sigma}^2_{B}$	31.257
$\hat{\sigma}_{TIMSS_{adv}} = \hat{B}^2\hat{\sigma}^2_{NAEP_{adv}} + \hat{\sigma}^2_{A} + 2\left(NAEP_{adv}\right)\hat{\sigma}_{AB} + \left(NAEP_{adv}\right)^2\hat{\sigma}^2_{B}$	43.946

The standard errors of linking reported in tables 8 and 9 are the square roots of the linking error variances in Tables 20A and 20B.

It is instructive to compare the standard error of linking for the projected NAEP mean to the standard error of linking for the projected NAEP achievement levels. Because the linking error is smaller at the mean, the standard error of linking for the NAEP projected achievement levels should be larger than for the mean. In fact, this is the case. The standard error of linking curves are presented in the following graphs. The standard error of linking for the projected mean of 498 in mathematics is 4.73 and for the projected mean of 510 in science are 5.43. In both cases, the standard error of linking for the mean is smaller than the standard error of linking for the achievement levels reported in tables 3 and 4.





One interesting question in linking studies is, "How much of the linking error is due to sampling and how much is due to test unreliability (or measurement error)?" In this study, we can answer that question by comparing the error variances in Tables 18A, 18B (sampling error in linking), and 19A, 19B (measurement error in linking), to Tables 20A and 20B (total error in linking). Tables 21A and 21B show the percent of linking error variance accounted for by sampling and measurement error.

Table 21A: Variance components of linking error for NAEP achievement levels projected on to the TIMSS grade 8 mathematics scale

	Sampling	Measurement
Basic	98.1%	1.9%
Proficient	96.4%	3.6%
Advanced	90.6%	9.4%

Table 21B: Variance components of linking error for NAEP achievement levels projected on to the TIMSS grade 8 science scale

	Sampling	Measurement
Basic	94.2%	5.8%
Proficient	93.8%	6.2%
Advanced	91.8%	8.2%

The main message of Tables 21A and 21B is that the vast majority of linking error is due to sampling. However, measurement error becomes a larger percentage of the linking error in the tails of the achievement distribution. This is why the measurement error for the advanced achievement level is a larger component of the linking error variance. The advanced achievement level is very high on the scale, where the measurement error is larger.

Another interesting question is "How much of the total survey error is due to linking error and sampling error"? The answer varies by country. Table 22A and 22B show the breakdown for the 2003 United States TIMSS.

Table 22A: Percent of total error variance due to linking and sampling for NAEP achievement levels projected on to the TIMSS grade 8 mathematics scale

	Linking Error	Sampling Error
Basic	59.3%	40.7%
Proficient	62.2%	37.8%
Advanced	73.8%	26.2%

Table 22B: Percent of total error variance due to linking and sampling for NAEP achievement levels projected on to the TIMSS grade 8 science scale

	Linking Error	Sampling Error
Basic	58.3%	41.7%
Proficient	59.6%	40.4%
Advanced	67.5%	32.5%

In Tables 22A and 22B we see that the linking error is always larger than the sampling error for all three achievement levels. For the Advanced level the linking error is two to three times the size of the sampling error. In other words the dominate source of error was due to linking, not sampling. Another way of saying this is that the error variance in this report is greater than the error variance in the 2003 TIMSS report. This is because the 2003 TIMSS does not have linking as a component of error, whereas linking is the major source of error in this report. The moral of this story is that there is *substantial* error in linking studies and that is why they should always be calculated, reported and taken into account in significance testing.

Linking error variance for the percent at and above projected achievement levels

So far in this technical appendix, all the error variances have been calculated in the scale score metric. However, the report is really about the percentages of students at and above various achievement levels (inverse cumulative percentages). Thus we must express the standard errors of linking in the inverse cumulative percentage metric as well as the scale score metric. This was done by making the assumption that the population distribution in each country is approximately normal. We know this assumption may not be true in some very low-performing and very high-performing countries. However, even in these circumstances, the normality assumption should still provide reasonable approximations. Suppose that the TIMSS achievement of students θ is normally distributed in country *j* with $\theta \sim N\left(\mu_j, \sigma_j\right)$. Estimates, $\hat{\mu}_j$ and $\hat{\sigma}_j$ of μ_j and σ_j are available from the published international reports of 1999 TIMSS and 2003 TIMSS. Let θ_C represent the cut-score on the TIMSS scale for the projected NAEP achievement level. Given the normality assumption, the percentage of students at and above each projected achievement level is

$$P_j(\theta > \theta_C) = \left[1 - \Phi\left(\frac{\theta_C - \hat{\mu}_j}{\hat{\sigma}_j}\right)\right] * 100, \tag{0.5}$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of a standard normal distribution.

However, we know that there is linking error (*LE*) in the projected achievement levels. Let $\theta_{C+\sigma_{LE}}$ be the upper limit of the margin of error interval for linking and $\theta_{C-\sigma_{LE}}$ be the lower limit. Then the percentage, P_j of students at and above the achievement level θ_C is between the upper and lower limit of the margin of error interval. The upper and lower limits are

$$\begin{aligned} P_j\left(\theta > \theta_{C+\sigma_{LE}}\right) &= \left[1 - \Phi\left(\frac{\theta_{C+\sigma_{LE}} - \hat{\mu}_j}{\hat{\sigma}_j}\right)\right] * 100, \text{ and} \\ P_j\left(\theta > \theta_{C-\sigma_{LE}}\right) &= \left[1 - \Phi\left(\frac{\theta_{C-\sigma_{LE}} - \hat{\mu}_j}{\hat{\sigma}_j}\right)\right] * 100, \end{aligned} \tag{0.6}$$

Although the upper and lower limits of the margin of error $P\left(\theta > \theta_{C+\sigma_{LE}}\right)$ and $P\left(\theta > \theta_{C-\sigma_{LE}}\right)$ are asymmetrical around P_j, a rough standard error of linking in the inverse cumulative percent metric can be obtained by

$$\sigma_{LEj} = \frac{P\left(\theta > \theta_{C-\sigma_{LE}}\right) - P\left(\theta > \theta_{C+\sigma_{LE}}\right)}{2} \tag{0.7}$$

Sampling error variance for the percent at and above projected achievement levels[13]

[13] The standard errors for sampling reported in this paper are more accurately estimated (and usually smaller) than those reported in Phillips (2007). I want to thank Tao Jiang for working out the statistical procedures for accomplishing this task.

Because TIMSS is a survey that is administered in each country, all statistics derived from it will have sampling error. Therefore, the percent of students at and above each projected achievement level P_j will have sampling error associated with it in equation (0.5). The sampling error can be estimated from the published international reports by calculating the standard error of a percentage

$$\sigma_{SEj} = \sqrt{\frac{P_j\left(100 - P_j\right)}{eff\left(n_j\right)}}\,. \qquad (0.8)$$

The quantity $eff\left(n_j\right)$ is the effective sample size associated with P_j (i.e., the actual sample size of the TIMSS survey divided by the design effect for P_j).

Total error variance for the percent at and above projected achievement levels

The total standard error for the percent of student at and above each achievement level P_j is the square root of the sum of the squared linking error (0.7) and squared sampling error (0.8).

$$\sigma_{Ej} = \sqrt{\sigma_{LEj}^2 + \sigma_{SEj}^2} \qquad (0.9)$$

The standard errors for projected achievement levels are reported in Tables 23 and 24.

Table 23: Percent At and Above proficient Projected on 2003 TIMSS Mathematics

Country	Percent	Standard Error
Singapore	73	2.7
Hong Kong, SAR	66	3.1
Korea, Rep. of	65	2.5
Chinese Taipei	61	2.7
Japan	57	2.8
Belgium (Flemish)	40	3.1
Netherlands	38	3.6
Hungary	37	2.9
Estonia	36	3.3
Slovak Republic	28	2.6
Australia	27	2.9
Russian Federation	26	2.8
Malaysia	26	2.9
United States (TIMSS)	26	2.5
Latvia	25	2.7
Lithuania	24	2.3
Israel	24	2.3
England	22	2.8
Scotland	22	2.6
New Zealand	21	3.0
Sweden	21	2.4
Serbia	19	1.8
Slovenia	19	2.2
Romania	18	2.2
Armenia	18	1.9
Italy	17	2.1
Bulgaria	17	2.1
Moldova, Rep. of	12	1.7
Cyprus	11	1.4
Norway	9	1.4
Macedonia, Rep. of	8	1.2
Jordan	7	1.1
Egypt	5	0.8
Indonesia	5	1.0
Palestinian Nat'l Auth.	4	0.6
Lebanon	3	0.7
Iran, Islamic Rep. of	2	0.5
Chile	2	0.4
Bahrain	2	0.4
Philippines	2	0.6
Tunisia	1	0.2
Morocco	1	0.2
Botswana	0	0.1
South Africa	0	0.2
Saudi Arabia	0	0.1
Ghana	0	0.1

Table 24: Percent At and Above proficient Projected on 2003 TIMSS Science

Country	Percent	Standard Error
Singapore	55	3.1
Chinese Taipei	52	3.4
Korea, Rep. of	45	3.3
Hong Kong, SAR	44	3.8
Japan	42	3.2
Estonia	41	3.7
England	38	3.5
Hungary	38	3.2
United States (TIMSS)	31	2.8
Netherlands	31	3.7
Australia	30	3.2
Sweden	28	2.9
New Zealand	26	3.5
Slovak Republic	26	2.8
Lithuania	25	2.7
Slovenia	24	2.8
Russian Federation	24	2.9
Scotland	24	2.7
Belgium (Flemish)	22	2.8
Latvia	21	2.7
Malaysia	20	2.9
Israel	18	2.0
Bulgaria	17	2.2
Italy	17	2.1
Jordan	15	1.9
Norway	15	2.0
Romania	14	2.0
Serbia	12	1.5
Macedonia, Rep. of	10	1.4
Moldova, Rep. of	10	1.6
Armenia	10	1.5
Egypt	8	1.1
Palestinian Nat'l Auth.	8	1.1
Iran, Islamic Rep. of	6	1.0
Cyprus	6	0.9
Bahrain	4	0.8
Chile	3	0.6
Indonesia	3	0.7
Philippines	3	0.7
Lebanon	3	0.6
Saudi Arabia	1	0.3
Botswana	1	0.2
South Africa	1	0.3
Morocco	1	0.2
Ghana	0	0.2
Tunisia	0	0.1

Technical Appendix B: Significance Testing and Multiple Comparisons

If we only conducted one significance test between country *A* and country *B* then a 95% confidence interval would be $95\%CI = \pm Z_{\alpha/2}\sqrt{\sigma^2_{E(A)} + \sigma^2_{E(B)}}$. However, when conducting a large number of hypotheses testing an adjustment for α is often used to compensate for the fact that many significance tests are being performed. If we have *k* independent tests, each at level α , then the probability that at least one is falsely rejected is $1-(1-\alpha)^k = \alpha_k$. For example, in the state-by-nation comparisons for mathematics there are 46 comparisons each state may wish to make (46 national comparisons with each state). With each $\alpha = .025$ (i.e., $\alpha = .05$ with a 2-tailed test), the family-wise error rate is $\alpha_k = .69$, so the probability of a false positive (or type-I error) among the 46 comparisons is equal to .69. When conducting multiple hypothesis tests we usually want to control α_k . This is referred to as controlling the family-wise error rate. The most common type of control for the family-wise error rate is the Bonferroni procedure (Bonferroni, 1936) where the α for each test would be $\alpha = \frac{\alpha_k}{46} = \frac{.025}{46} = .000543$. With this procedure you divide the significance level for each test by the number of significance tests so that the family-wise error rate is $\alpha_k = \frac{\alpha}{k}$, therefore $\alpha_k = 1-\left(1-\frac{.025}{46}\right)^{46} = .025$. Unfortunately, the Bonferroni procedure suffers from low power properties when the number of tested hypotheses is large.

False Discovery Rate (FDR): Instead of controlling for the chance of *any* false positive (like the Bonferroni procedure), the FDR controls for the proportion of false positives (Benjamini, Y., and Hochberg, Y., 1994). *The FDR is the expected proportion of true null hypotheses rejected out of the total number of null hypotheses rejected*. Multiple comparison procedures controlling the FDR are more powerful than the commonly used multiple comparison procedures based on the family-wise error rate. FDR controlling procedures are especially suited to situations where there are a large number of hypotheses being tested. Suppose *k* hypotheses are tested, and *R* of them are rejected. Of the rejected hypotheses, suppose that *V* of them are really null (i.e., V is the number of type I errors, or false positives). The False Discovery Rate is defined as $E\left(\frac{V}{R}\right)$, where *E* is the expected value. Let $H_1 ... H_k$ be the null hypotheses and $P_1 ... P_k$ their corresponding p-values. The p-values have been ordered from lowest (most significant) to highest (least significant). For each P_j we calculate Q_j where $Q_j = \frac{j}{k}\alpha$. If $P_j \le Q_j$, we reject the null hypothesis. The FDR is used in this paper for all significance testing.