This product is part of the RAND Corporation technical report series. Reports may include research findings on a specific topic that is limited in scope; present discussions of the methodology employed in research; provide literature reviews, survey instruments, modeling exercises, guidelines for practitioners and research professionals, and supporting documentation; or deliver preliminary findings. All RAND reports undergo rigorous peer review to ensure that they meet high standards for research quality and objectivity.

TECHNICAL REPORT

Lessons from the Field

Developing and Implementing the Qatar Student Assessment System, 2002–2006

Gabriella Gonzalez • Vi-Nhuan Le • Markus Broer • Louis T. Mariano
J. Enrique Froemel • Charles A. Goldman • Julie DaVanzo

Prepared for the Supreme Education Council

RAND RAND-QATAR POLICY INSTITUTE

The research described in this report was prepared for the Supreme Education Council and conducted within the RAND-Qatar Policy Institute and RAND Education, programs of the RAND Corporation.

Library of Congress Cataloging-in-Publication Data

Lessons from the field : developing and implementing the Qatar student assessment system, 2002-2006 / Gabriella Gonzalez ... [et al.].
p. cm.
Includes bibliographical references.
ISBN 978-0-8330-4689-5 (pbk. : alk. paper)
1. Educational tests and measurements—Qatar. 2. Students—Rating of—Qatar. I. Gonzalez, Gabriella.

LB3058.Q38L47 2009
371.26'2095363—dc22

2009009273

Published 2009 by the RAND Corporation
1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, VA 22202-5050
4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213-2665
RAND URL: http://www.rand.org
To order RAND documents or to obtain additional information, contact
Distribution Services: Telephone: (310) 451-7002;
Fax: (310) 451-6915; Email: order@rand.org

Preface

His Highness the Emir of Qatar sees education as the key to Qatar's economic and social progress. Long concerned that the country's education system was not producing high-quality outcomes and was rigid, outdated, and resistant to reform, the Emir approached the RAND Corporation in 2001, asking it to examine the kindergarten through grade 12 (K–12) education system in Qatar and to recommend options for building a world-class system consistent with other Qatari initiatives for social and political change. In November 2002, the State of Qatar enacted the Education for a New Era (ENE) reform initiative to establish a new K–12 education system in Qatar.

One component of ENE was the development of internationally benchmarked curriculum standards in modern standard Arabic, English as a foreign language, mathematics, and science subjects. These standards are used in the Independent schools that have been developed as part of the reform. Qatar also established a standardized, standards-based student assessment system to measure student learning vis-à-vis the new curriculum standards among all students in government-sponsored schools, including the Independent schools, the traditional Qatar Ministry of Education schools, and private Arabic schools, which follow the Qatar Ministry of Education curriculum in a private-school setting. The development of a comprehensive assessment system, its alignment with the standards, and its standardized administration to the targeted students are vital components of ensuring the success of Qatar's ENE reform. The system allows parents to gauge the performance of different schools and allows policymakers to monitor school quality.

From July 2002 to July 2005, RAND assisted in the implementation and support of the ENE reform. The reform design and the results of the first two years of implementation are reported in the RAND monograph *Education for a New Era: Design and Implementation of K–12 Education Reform in Qatar* (Brewer et al., 2007).

This technical report describes work carried out as part of the larger RAND study. It documents the development of the Qatar Student Assessment System (QSAS) with particular attention to its primary component, the Qatar Comprehensive Educational Assessment (QCEA), expanding on the discussion of the assessment system in Brewer et al. (2007). Staff of the Supreme Education Council's (SEC's) Evaluation Institute and the RAND Corporation collaborated on the QSAS design and implementation and jointly authored this report. (Coauthors Markus Broer and Juan Enrique Froemel have since left the Evaluation Institute.) This report should be of interest to education policymakers or test developers in other countries looking to develop standards-based assessments, as well as to researchers and practitioners interested in recent education reforms undertaken in Qatar and in the Middle East region in general.

More detailed information about the reform can be found at the SEC Web site: www.english.education.gov.qa (English version, with a link to the Arabic version).

This project was conducted under the auspices of the RAND-Qatar Policy Institute (RQPI) and RAND Education in conjunction with Qatar's Student Assessment Office. RQPI is a partnership of the RAND Corporation and the Qatar Foundation for Education, Science, and Community Development. The aim of RQPI is to offer the RAND style of rigorous and objective analysis to clients in the greater Middle East. In serving clients in the Middle East, RQPI draws on the full professional resources of the RAND Corporation. RAND Education analyzes education policy and practice and supports implementation of improvements at all levels of the education system.

For further information on RQPI, contact the director, Richard Darilek. He can be reached by email at Richard_Darilek@rand.org; by telephone at +974-492-7400; or by mail at P.O. Box 23644, Doha, Qatar. For more information about RAND Education, contact the associate director, Charles Goldman. He can be reached by email at Charles_Goldman@rand.org; by telephone at +1-310-393-0411, extension 6748; or by mail at the RAND Corporation, 1776 Main Street, Santa Monica, California 90401, USA.

Contents

Figures

Tables

Summary

Background

The Arabian Gulf nation of Qatar has recently positioned itself to be a leader in education reform. The country's leadership has initiated a number of changes to Qatar's kindergarten through grade 12 (K–12) and higher education systems. In 2001, the Emir of Qatar, His Highness Sheikh Hamad Bin Khalifa Al Thani, asked RAND to help redesign the country's K–12 education system. RAND recommended that Qatar institute a comprehensive education reform with a standards-based education system at its core. In 2002, implementation of the reform initiative, Education for a New Era (ENE), began.

ENE is based on four core principles: *variety* in educational offerings, *choice* for parents to select schooling options for their children, *autonomy* of newly opened schools, and *accountability* for all government-sponsored schools in Qatar, including newly developed Independent schools, traditional public schools operated by the Qatar Ministry of Education, and private Arabic schools that follow the Ministry of Education curriculum in a private-school setting.

Central to ENE was the development of internationally benchmarked curriculum standards in modern standard Arabic (*fusHa*), English as a foreign language, mathematics, and science for students in grades K–12. The curriculum standards include both content standards, which note what students should be taught in each grade, and performance standards, which note what students should know by the end of each grade. Curricula, assessments, and professional development are aligned with and follow from the curriculum standards. In the 2004–2005 academic year, 12 Independent schools opened and began operating alongside the traditional Ministry of Education schools. The Independent schools are governed by the Supreme Education Council (SEC), which was established as part of the reform plan. Independent schools follow the established curriculum standards, but principals of the schools have more autonomy to make decisions about educational approach (e.g., curricula used in the classrooms), staffing policies, and budget spending than do principals in Ministry of Education schools. More Independent schools have opened in each academic year, with 85 operating during the 2008–2009 school year. Ministry schools are still in operation, running in tandem with the Independent school system.

The SEC includes two new government institutes. The Education Institute developed the standards in 2005, funds and oversees the Independent schools, and provides professional development for teachers and staff in Ministry and Independent schools. The Evaluation Institute developed and administers the standards-based assessments as well as the student, parent, teacher, and school administrator surveys. School-level results from the surveys and assessments are reported on publicly available school report cards. Parents can use the school report

cards to inform their decisionmaking on where to send their children to school. Starting in 2006, individual- and classroom-level reports are provided to parents and teachers, respectively. Parents can use the individual reports to follow their children's progress from year to year, and teachers can use the classroom reports to help guide their teaching.

Building the Qatar Student Assessment System

From 2002 through 2005, RAND assisted the SEC with the implementation of the early stages of the reform. In that time, RAND and the Evaluation Institute's Student Assessment Office (SAO) crafted a design for Qatar's standards-based student assessment system, the Qatar Student Assessment System (QSAS). The design called for the QSAS to provide (1) information about school performance to the public to motivate school improvement and promote informed parental choice; (2) feedback to teachers, helping them tailor instruction to support the needs of student bodies; and (3) detailed information to policymakers about the education reform's progress in general and, specifically, about Independent schools' performance for accountability purposes.

To serve these three purposes, the initial design of the QSAS included multiple types of standardized and systematic assessments, each measuring the learning and achievement of students in a variety of skills and competencies described in the newly developed curriculum standards. Examples of such assessments included a large-scale summative assessment administered at the end of the school year, performance assessments (such as hands-on science experiments) that would be evaluated by a team of local experts, and in-class, computer-delivered formative assessments administered throughout the school year. The results of the assessments could be tracked in a database managed by the Evaluation Institute.

In the first years of the reform, RAND and the SAO focused on the development of one component of the QSAS—the Qatar Comprehensive Educational Assessment (QCEA). The QCEA is the first national, standardized, standards-based assessment in the region. The QCEA measures student learning and performance according to the requirements set forth in the curriculum standards using a multiple-choice and open-ended question format. It is a summative assessment and is administered at the end of the school year.

The development of the QSAS and QCEA involved contractors and experts from around the world: Europe, the Middle East, South America, and the United States. Through the QCEA development, implementation, and process to align its questions with the Qatar curriculum standards, the SAO and RAND worked closely with test developers Educational Testing Service (ETS) and CTB/McGraw-Hill (CTB); the curriculum standards-development contractor, the Centre for British Teachers (CfBT, now the CfBT Education Trust); and the contractor charged with assisting in the development of the national educational surveys and administration of the surveys and assessments, the National Opinion Research Center (NORC).

The first administration of the QCEA occurred in April and May 2004, before the opening of the Independent schools or the finalization of the new curriculum standards, to students in grades 1–12. The 2004 test provided a snapshot of student achievement vis-à-vis general standards to measure what a student is expected to do or know in mathematics, science, English as a foreign language, and Arabic. In 2005, the QCEA was revised to align it with the curriculum standards. In 2004, the results of the QCEA were reported as percent correct. In 2005 and 2006, it was administered to students in all government-sponsored schools in

grades 4–11. (In 2005, math, English, and Arabic assessments were given to students in grades 1–3.) Starting in 2007, the QCEA was administered only to students in the Independent schools. From 2005 onward, the QCEA reported performance levels, with students measured according to five levels: meeting standards, approaching standards, below standards–may approach standards with some additional effort, below standards–may approach standards with considerable additional effort, and below standards–may approach standards with extensive additional effort.

In each year from 2004 through 2006, the QCEA was fielded to about 88,000 students in Ministry, private Arabic, and Independent schools—approximately 95 percent of the target population. Qatar now has the tools at its disposal to understand the educational achievement of its student population and inform policymaking. Prior to these reform efforts, little systematic, objective information on student achievement and skills existed. Although a number of changes have been made to the testing operation since its inception, and a number of improvements to the QSAS can still occur, the advent of the QCEA has forever changed the educational landscape of the country.

Purpose and Approach of This Report

This report documents the initial design of the QSAS and chronicles the development and administration of the QCEA. The work reported here was carried out jointly by RAND and the SAO. In this report, we draw lessons for future assessment development in Qatar and for education policymakers in other countries considering a standards-based approach to student assessment.

In writing this report, we relied on three sources of information. First, to contextualize the design of the QSAS and QCEA, we reviewed the fields of accountability, standards-based education, assessment theory, and practitioners' guides to developing assessments. Second, to elaborate on the decisionmaking process for key policies, we reviewed the minutes of meetings held between July 2002 and July 2005 among representatives from RAND, the SAO, the Evaluation and Education Institutes, and the contractors that assisted in the development and administration of the assessments. Third, to further explain decisionmaking processes, we reviewed internal memos—from both RAND and the SAO.

Limitations of This Report

Given the historical nature of this report, it is important to keep in mind several limitations. First, this report is limited in scope. It is not meant to be a testing technical report, nor do we assess the validity of the results of the tests to serve the hoped-for purposes. Although valuable and a necessary part of any testing effort, such an analysis is beyond this report's scope. A second limitation is that it provides only the perspective of the RAND and SAO teams and not those of the other Evaluation and Education Institute staff and contractors with whom we worked in aligning the assessments with Qatar's curriculum standards and in administering those assessments. A third limitation is that it was difficult, at times, to uncover who within the governance structure of the reform effort made certain decisions about the assessment system, so we are not always able to attribute decisions.

Lessons Learned

A number of important lessons emerged from our experience that can be useful to education policymakers in Qatar as they move the QSAS forward and to education leaders around the world considering implementing a standards-based assessment system. These are summarized in the remainder of this section.

The separation of standards development and assessment development in two offices hampered communication in terms of alignment. The design of the reform effort placed responsibility for developing the standards with one entity, the Curriculum Standards Office (CSO) within the Education Institute, and responsibility for developing the assessments with another, the SAO within the Evaluation Institute. Although few informal linkages developed, these proved too tenuous to encourage cross-office discussions. We recommend that, prior to implementation, formal linkages between standards-development and assessment-development authorities be built. One option to improve the alignment process is to have a permanent staff member with explicit duties to liaison between the two offices. Alternatively, the curriculum staff and assessment-development staff can be housed within the same office.

The timeline for developing a fully aligned standards-based assessment system was too short. The education leadership in Qatar expected to have a standards-based assessment system in place by the end of the 2004–2005 academic year—the first year that Independent schools were open. The SAO, RAND, and the test developers encountered a number of challenges in meeting this deadline: By 2005, the QSAS's goals, purposes, uses, and design features were laid out, but the SAO and RAND were unable to finalize a detailed blueprint or implement the system's features by this date. There were three reasons for this delay. First, given the tight timeline, the SAO and RAND decided to focus efforts on developing the core component of the QSAS, the QCEA, as it was to be the largest and most comprehensive component of the system. Second, in 2003 and 2004, the SAO had only three staff members, which limited the office's capacity to focus on the implementation of the QCEA alongside the implementation of other components of the QSAS. Third, the SAO, the test developers, and RAND worked with draft curriculum standards until they were finalized in 2005. Therefore, final decisions about the QSAS design could not occur until the standards were finalized. To allow for appropriate time to develop, pilot, and field a fully aligned, comprehensive assessment system, we recommend a minimum of three years, as suggested by experts (Commission on Instructionally Supportive Assessment, 2001; Pellegrino, Chudowsky, and Glaser, 2001), with even more time if performance-based assessments are to be applied. For education systems that may encounter similar staff challenges and the possibility of rapid policy shifts, as experienced in Qatar, we recommend five years.

Logistic and administrative constraints often took precedence over the substantive needs of the QCEA testing operation. In the first year of the QCEA, the Evaluation Institute made a number of operational decisions that prioritized logistical issues over substantive issues as a way to ease the perceived burden on test administrators and students. For example, for the pilot test of the QCEA in 2004, the length of test time was limited to one class period so as not to disturb the classroom schedule. However, the test developers noted that the amount of test time was inadequate—particularly for the mathematics tests, for which students were expected to use tools and other manipulatives when answering the questions. Test time was subsequently lengthened to accommodate the test's psychometric requirements and to ensure that the test was as fully aligned with the standards as possible. The prioritization of logistics may have

occurred because members of the Evaluation Institute in charge of test administration had no experience with delivering, coding, or managing a testing operation of the size and scope of the QCEA. We recommend that, prior to the administration of a test, the entities in charge of developing and administering the tests agree on administration processes and procedures that strike a balance between limiting student burden or fatigue and ensuring that appropriate analyses can be made from the tests' results.

Many testing policies did not consider existing research or analysis. A number of policies concerning the testing operation did not consider available research, which, in turn, confused schools and may have had potentially negative long-term effects. One example of this was having Independent schools move toward teaching mathematics and science in English and the subsequent decision to offer mathematics and science QCEA tests in English for schools that chose this option. These decisions were made without considering Evaluation Institute studies on whether this would be a helpful policy for the students, who may have trouble mastering mathematics and science content in a second language. We therefore recommend that, in making decisions, education policymakers consider research findings and empirical evidence. If the Evaluation Institute, the Education Institute, and the governing body of the SEC are to make informed policy decisions about the assessments and student achievement, they must base those decisions on empirical evidence, lest innuendo or unfounded perceptions sway education policy in the nation.

There was insufficient communication about the purposes and uses of testing. Understandably, the public had many questions about the purpose of the QSAS and, in particular, the QCEA and its implications for students in Qatar's schools. Yet, the SEC and the Evaluation Institute provided little public information to answer these questions. The QSAS communication effort can be improved by incorporating direct outreach efforts:

- Outreach programs for parents and other community stakeholders might be scheduled for weekends or weeknights, when working adults can attend meetings. (For Qataris, evening meetings would be the most appropriate option.)
- Outreach for education stakeholders should occur on a continuous basis throughout the early years of testing. (For Qatar, these stakeholders include Independent school operators, teachers, and Ministry of Education personnel.)

Furthermore, public acceptance of the assessment system could have been enhanced by improving the transparency of the testing operation. In other testing operations, this problem could be addressed early on by providing individual-level achievement data from the first year of testing. (For the QCEA, individual-level data were available only after the third year of testing.)

Challenges to Address in the Future

The QSAS is still in its nascent stages, and a number of challenges still exist for the Evaluation Institute:

- The standards for secondary school students are divided into foundation and advanced levels. The QCEA now tests foundation standards only. Future versions of the QCEA

will have to consider testing the advanced standards as more students start to learn those standards.

- Students with learning or developmental disabilities are not presently included in the testing operation but tend to be mainstreamed with traditional students in Qatar. To incorporate these students into the QSAS, the Evaluation Institute will need to develop testing accommodations for those with disabilities.
- At some point, the Education Institute will modify the Qatar curriculum standards. The Evaluation Institute needs to be prepared to make continuous appraisals of how well the QCEA aligns with the standards and make any adjustments to the test battery if changes to the standards occur.
- A number of the standards could be tested appropriately with the use of a computer. In its quest to assess student learning of the standards, the Evaluation Institute should explore how best to incorporate computer technology in the testing operation and whether computer-based delivery of assessments is feasible given the country's information technology infrastructure.
- Parents continue to have questions about the QSAS and, specifically, doubt whether it is necessary. To promote public acceptance, the Evaluation Institute will need to enhance communication with the public so that QCEA results can inform parental choice, school accountability, and educational policymaking. This should include reports of interest to practitioners and studies to test the validity of using QCEA results to inform school- or classroom-level educational decisions.
- Short- and long-term ramifications of a recent decision to limit the testing operation to students in the Independent schools will have to be carefully weighed against the goals and principles of the reform effort.

Acknowledgments

We thank the Emir of Qatar, His Highness Sheikh Hamad Bin Khalifa Al Thani, and his Consort, Her Highness Sheikha Mozah Bint Nasser Al Missned, for initiating the improvement of education in Qatar. We also thank members of the SEC's Executive Committee, Sheikha Abdulla Al Misnad and Mohammed Saleh Al Sada, for their continued support of the reform effort.

We also acknowledge the efforts of members of the Evaluation Institute who were instrumental in the design and application of the QCEA, including the director of the Evaluation Institute, Adel Al Sayed; staff of the SAO, Mariam M. Abdallah Ahmad, Sharifa Al Muftah, Huda Buslama, Asaad Tournatki, and Abdesalam Buslama; and staff of the Data Collection and Management Office, Salem Al Naemi, Jamal Abdulla Al Medfa, and Nasser Al Naemi. We also acknowledge key staff at ETS, Lynn Zaback, Jenny Hopkins, Mary Fowles, and Paul Ramsey; CTB, Robert Sanchez, Gina Bickley, William Lorie, and Diane Lotfi; and NORC, Craig Coelen, Hathem Ghafir, and Eloise Parker.

This report benefited from reviews by Susan Bodilly, Cathleen Stasz, Brian Stecher, and Derek Briggs. Paul Steinberg deftly assisted in organizing an early draft of the document. The authors alone are responsible for the content and any errors herein.

Abbreviations

AM	alignment meeting
CAT	computer-adaptive testing
CfBT	Centre for British Teachers
CSO	Curriculum Standards Office
CTB	CTB/McGraw-Hill
DCMO	Data Collection and Management Office
ENE	Education for a New Era
ETS	Educational Testing Service
GCE	General Certificate of Education
HEI	Higher Education Institute
IELTS	International English Language Testing System
IB	International Baccalaureate
K–12	kindergarten through grade 12
NORC	National Opinion Research Center
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
QCEA	Qatar Comprehensive Educational Assessment
QNEDS	Qatar National Educational Database System
QSAS	Qatar Student Assessment System
RQPI	RAND-Qatar Policy Institute
SAO	Student Assessment Office
SEC	Supreme Education Council
SEO	School Evaluation Office

TIMSS	Trends in International Mathematics and Science Study
TOEFL	Test of English as a Foreign Language

Glossary

The following terms are defined within the context of educational assessment.

Bookmark method. A method used to set cut scores to determine performance levels for assessment results, created by CTB/McGraw-Hill in 1996. Using item response theory, test questions are ordered on a scale of difficulty, from easy to hard, and are presented in this order to a panel of experts. Each panel member places bookmarks in the booklet of reordered test items at points that, in his or her opinion, correspond best to the performance descriptions. Bookmark placements are averaged and the results of the decisions (percentage of students in each performance category) are then discussed.

Computer-adaptive testing (CAT). An assessment in which questions are administered to the examinee according to his or her demonstrated proficiency in "real time." Based on answers to previous items, a computer-adaptive test presents either harder or easier test questions that better fit the proficiency level of the examinee.

Computer-delivered testing. An assessment that is administered to the examinee by computer. The test may or may not be computer-adaptive.

Constructed-response item. An open-ended question on an assessment to which the examinee writes his or her own response.

Curriculum standards. Descriptions of skills, content, and competencies that a student must learn and be able to demonstrate, by subject and grade level.

Diagnostic assessment. An assessment of a student's strengths and weaknesses that is administered before the student begins a particular learning task or series of tasks and that guides what types, intensity, and duration of interventions might be needed.

Depth of knowledge. A term that refers to the different complexity levels that items or curricular objectives demand. For example, a lower level may be assigned to a recall item, while a higher-level item might require more complex reasoning skills. Depth-of-knowledge consistency is one of the criteria used for judging the alignment between the Qatar Curriculum Standards and the QCEA.

Formative assessment. A test that gathers information about learning as learning is taking place. Teachers use formative assessments to improve student learning; such assessments often take the form of in-class work or homework.

General Certificate of Education (GCE). A secondary-level academic certification system used in Britain and in some former British colonies. It is often divided into two levels: ordinary level (O-level) and advanced level (A-level), although other categories exist. Since 1999, the advanced subsidiary level (AS-level) has also come into wider use. In 1986, O-level qualifications were replaced by a new system, the General Certificate of Secondary Education.

International Baccalaureate (IB). An educational foundation established in 1968 in Geneva, Switzerland. As of October 2008, the IB organization works with 2,405 schools in 131 countries to develop and offer three curricular programs to more than 658,000 students age 3 to 19 years. The Primary Years Programme is for students age 3–12, the Middle Years Programme is for students age 11–16, and the Diploma Programme is for students age 16–19.

International English Language Testing System (IELTS). A test of "international English" language proficiency that includes British English and American English (in contrast to the Test of English as a Foreign Language (TOEFL), which focuses on North American English). The IELTS tests the ability to speak, read, write, and listen to English and is required by many English-speaking universities and colleges outside of the United States.

Item. A question on an assessment.

Item response theory model. A psychometric model that describes the probability of an examinee's response on an assessment item as a function of his or her underlying proficiency and characteristics of the item. Item responses may be scored as right or wrong or on a more general ordinal categorical scale. Model parameters quantify the proficiency of each examinee and the characteristics of each item. Item characteristics typically describe the difficulty of the item and degree to which an item can discriminate among varying levels of proficiency. For multiple-choice items, a guessing parameter may be included to take into account that even students with very low proficiency may get some items right merely by guessing.

Modified Angoff method. A method used to set cutoff points, or cut scores, to determine performance levels for assessment results. A panel of experts determines the probability that a minimally competent student can answer each question on the test. These probabilities are then used to determine cut scores for the performance levels.

Multiple-choice item. A question on an assessment in which the examinee must choose one correct answer among a number of possible answers presented.

Paper-and-pencil level test. A type of test that consists of different forms (e.g., low, medium, and high) with content that is more closely matched to an individual's proficiency level. Ideally, the three forms overlap, sharing a common measurement range and some test items. Each deals with the same concepts and topics but at differing levels of complexity.

Performance level. A term describing a specific level of competence on an assessment. Performance levels for the QCEA are "meets standards," "approaches standards," and three levels of "below standards." Cut scores for these performance levels were determined by a panel of experts using the modified Angoff method for English and Arabic tests and the bookmark method for mathematics and science tests.

Performance-based assessment. An assessment that requires that a student perform a task, such as a scientific experiment, or generate an extended response, such as a research paper.

Pilot study. A field test of assessment items used to gain information on item performance to develop test forms for the main application of the test.

Portfolio. A collection of a student's work that typically shows his or her progress through a school year or term. Often, a panel of teachers judges the work to standardize the evaluation of the student's performance.

Programme for International Student Assessment (PISA). An internationally comparative paper-and-pencil and computer-delivered assessment that tests 15-year-olds' capabilities in reading literacy, mathematics literacy, and science literacy and is administered every three years. PISA emphasizes functional skills that students have acquired as they near the end of

mandatory schooling and assesses how well prepared students are for life beyond the classroom by focusing on the application of knowledge and skills in everyday situations. Students also complete a questionnaire to gauge their familiarity with information technology. Parents also complete a questionnaire.

Progress in International Reading Literacy Study (PIRLS). An internationally comparative assessment of reading literacy administered to fourth-grade students in their native language in more than 40 countries. This grade level was chosen because it is an important transition point in children's development as readers. Typically, at this point, students have learned how to read and are now reading to learn. Moreover, PIRLS investigates the impact of the home environment on reading; the organization, time, and materials for learning to read in schools; and the curriculum and classroom approaches to reading instruction.

Reliability. A term used to describe the degree to which items measure a common underlying construct in a test accurately (internal consistency) or the degree to which tests yield similar results over time (stability).

Summative assessment. A test that gathers information about learning after the learning has occurred, usually for the purpose of assigning grades to students.

TerraNova. The name of a series of standardized tests developed by CTB/McGraw-Hill.

Test of English as a Foreign Language (TOEFL). A test that evaluates the potential success of an individual to use and understand standard American English at the college level. It tests the ability to speak, read, write, and listen to English and is required for non-native English-speaking applicants at many colleges and universities in the United States and in other English-speaking countries.

Trends in International Mathematics and Science Study (TIMSS). An internationally comparative curriculum-based assessment of fourth- and eighth-grade students' mathematics and science achievement that is conducted every four years. TIMSS assessments offer a variety of multiple-choice and extended free-response items, requiring written explanations from students. Additional information from teacher, student, and school questionnaires provides a context for the achievement data and helps explain differences in achievement.

Usability study. A field test of assessment items used to evaluate basic item quality measures; not an official pilot test of items.

Validity. A term used to describe the degree to which a test measures the construct it purports to measure and the extent to which inferences made and actions taken on the basis of test scores are appropriate and accurate.

CHAPTER ONE

Introduction

Background on Qatar's Education System

Qatar is a geographically small country located on a peninsula off Saudi Arabia that extends into the Arabian Gulf. It is one of the wealthiest nations in the world because of its oil production and vast reserves of natural gas, coupled with a small citizen population of about 200,000. Before oil was discovered in 1947, no formal education system existed in Qatar. Instead, some children in villages and towns memorized passages from the Qur'an and learned to read and write in *kuttabs*—informal classes taught in mosques or homes by literate men and women who were knowledgeable about Islam. From these early days, the development of education in Qatar focused mainly on the male population. The Qatar Ministry of Education was established in 1956, ushering in an era of free education for both boys and girls. Public schooling is free to all Qatari schoolchildren and to expatriate children whose parents are employed by the government (Brewer et al., 2007).

Following independence from Britain in 1971, Qatar launched a period of educational development to match the demands and challenges of independence. These reform efforts centered on Qatar developing its identity as a sovereign state: Curriculum was developed "in house," and teacher-training programs were established to encourage Qataris to become teachers (Jolo, 2004). In 1973, Qatar's sole postsecondary education option was a teacher-training program with 150 students. In 1977, Qatar's only state-sponsored academically oriented university, Qatar University, was established. With these and more recent investments in education, the literacy rate among Qataris increased through the years, reaching 98.2 percent among 15- to 19-year-olds by 2004 (Qatar Planning Council, 2005).

Qatar's Ministry schools are divided into three levels: primary (grades 1–6), preparatory (grades 7–9), and secondary (grades 10–12). Girls and boys attend separate schools, and children are taught by teachers of the same gender as themselves.[1] In addition to the publicly funded government schools, a significant number of private schools serve both Qataris and citizens of other countries residing in Qatar. There are three types of private schools. One is "private Arabic" schools, which charge tuition and are geared toward Qataris and other Arabs who want to follow the Ministry curriculum but in a private-school setting. The second is "community" schools that cater to students from specific countries, are affiliated with a particular embassy, and use the curriculum of the country with which they are affiliated (e.g., the

[1] One exception to this rule is in "model schools" for boys in grades 1–4. These schools were developed to ease the transition for young boys from home to school, as well as to provide more employment opportunities for female teachers. In these schools, both the teaching staff and the administration are female. The first three model schools opened in 1978, and their success led to a five-year plan to implement this type of school system-wide (Brewer et al., 2007).

Pakistan Education Center follows Pakistan's national curriculum). The third type is "international" schools, which follow the curriculum of a country or an international curriculum but are not affiliated with a particular embassy and are open to students of many nationalities (e.g., Qatar Academy follows the International Baccalaureate, or IB, curriculum). Tuition rates vary widely depending on the type of private school.

The Context for Reforming Qatar's K–12 Education System

In 2001, the Emir of Qatar, His Highness Sheikh Hamad Bin Khalifa Al Thani, asked RAND to help redesign the K–12 education system. This led to the Education for a New Era (ENE) reform initiative, established by law in November 2002 by Emiri Decree No. 37 (Qatar Supreme Education Council, 2006). The overall goals of ENE were to improve student outcomes (broadly defined), enhance students' problem-solving and critical-thinking skills, socialize students to take a more active role in their communities and civic culture, and position Qatar as a world leader in education.

Qatar initiated the ENE reform to tackle perceived deficits in the quality of the K–12 education offered to its students: Prior to ENE, many of Qatar's students were retained each year, after-school tutoring was prolific because parents did not feel that their children were adequately learning in Ministry schools, and most secondary school graduates were unprepared to enter selective postsecondary institutions or science- and technology-related jobs. This lack of quality resulted from a number of problems inherent in the education system as a whole (Brewer et al., 2007):

- The Ministry of Education lacked a vision to implement its goals or initiate change. Instead, it reacted to problems as they arose, adding departments or processes in a piecemeal fashion rather than with a coherent vision in mind.
- The Ministry's hierarchical organizational structure did not foster innovation or change. Ironically, although the Ministry was very structured, parents, teachers, and other stakeholders did not know to whom to address complaints or suggestions because the lines of authority were unclear. Likewise, there appeared to be little effort from the Ministry to reach out to its stakeholder population and understand its needs.
- Students were taught an outdated and rigid curriculum, and teachers had to follow Ministry-mandated lesson plans each day. In addition, there were too many subjects to cover in the time allotted, resulting in superficial content coverage.
- With the focus on lecturing, few opportunities existed for student-teacher interaction in the classroom. The lecture style also did not allow teachers to customize their approaches for students with different abilities; learning in the Ministry schools was based on rote memorization.
- School administrators had little authority or flexibility. The Ministry assigned principals to buildings, assigned teachers and other staff to schools, and provided furniture, equipment, textbooks, and all other instructional materials.

- Finally, although teachers were held accountable for executing the centralized curriculum, no one was held accountable for students' performance.[2] There were no system-level goals for student outcomes; teachers and administrators had no sense of whether they were increasing students' knowledge or improving their skills.

Overview of the Education for a New Era Reform

To address the problems of the Ministry system and improve the rigor and quality of Qatar's education system with the goal of preparing Qatari graduates to contribute to and participate in a globalized economy and an increasingly democratic state, Qatar's leadership elected to pursue comprehensive education reform rather than target one component of the education system. RAND recommended a K–12 standards-based education system in which internationally benchmarked standards would be developed. Curriculum materials, assessments, and professional development were to be aligned with these standards. As part of this reform, curriculum standards were developed for the core academic subjects of mathematics, science, Arabic, and English as a foreign language for each grade level, from kindergarten through grade 12. The Qatari leadership chose these four subjects because they represented content areas that would help the nation compete in a global economy.[3] The curriculum standards for each grade level specify a challenging set of knowledge and skills that all students in Qatar's government-sponsored schools should possess and be able to demonstrate at the completion of that grade level. To promote continuous improvement to the system and to institute feedback loops in information and dissemination, the reform effort called for education data to be collected, analyzed, and disseminated to the public (Brewer et al., 2007).

ENE is based on four key principles:

- Promote the *autonomy* of education providers (teachers and school administrators).
- Provide a *variety* of government-funded schooling options from which parents can choose.
- Hold schools *accountable* to parents and the community for the education of the student body.
- Offer parents *choices* in terms of where to send their children to school.

---

[2] Educational testing in grades 1–12 in Qatar consisted of school-specific midyear and end-of-year tests administered at the preparatory and secondary grade levels and a national exam administered midyear and at the end of 12th grade. In this system, which is still in effect for Ministry and private Arabic schools, results from the two 12th-grade tests are added together and students receive a percent-correct score that is placed on a graduation certificate. Students who fail the tests are given another test over the summer. The two 12th-grade tests, known collectively as the National Exit Exam, assess student knowledge in the subjects associated with the curricular track that the student has followed in secondary school (humanities, science, or humanities and science). A group of Ministry of Education administrators develops a different National Exit Exam each year, but in 2005 and 2006, teachers were also asked to submit questions. Students who pass the tests receive a certificate of graduation, which makes them eligible to apply to a number of universities in the region. The score determines a student's eligibility for scholarships to study abroad and, until recently, entrance to Qatar University and placement in a job through the Qatar Ministry of Civil Service and Housing.

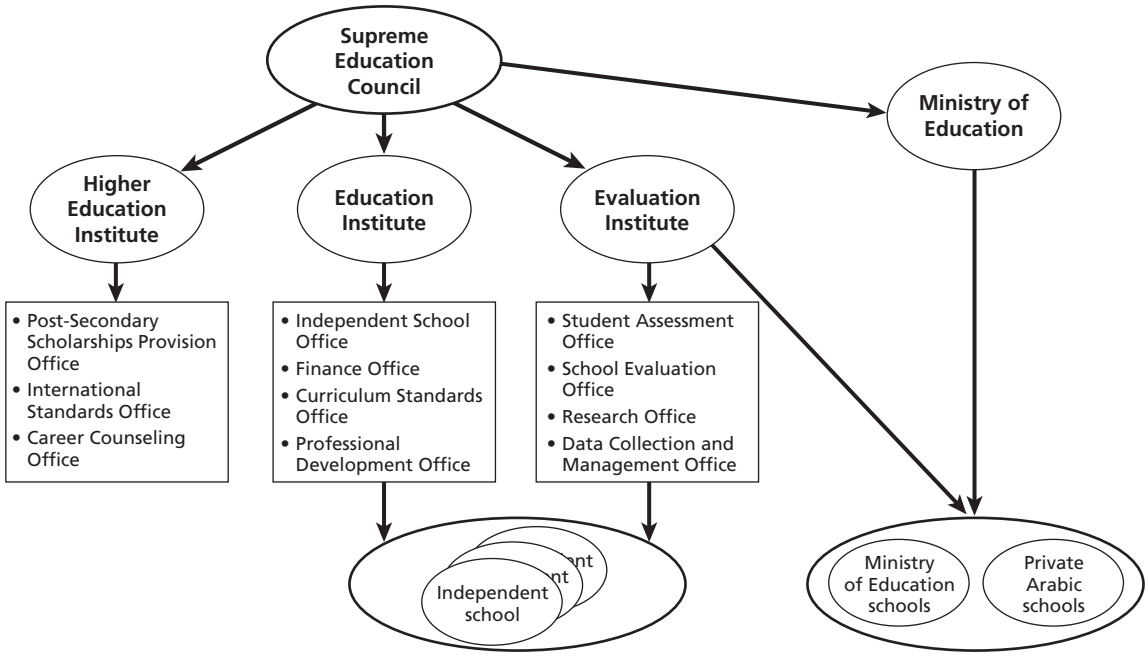
[3] Religious study, or shari'ah, was also considered important for children's education in Qatar. However, national curriculum standards are still in development. Instead, schools have been encouraged to use the shari'ah curriculum already in place in the Ministry of Education. For more information about the development of the new Qatar curriculum standards, see Brewer et al. (2007).

Governance Structure of the Education for a New Era Reform

ENE incorporated a new governance structure, the Supreme Education Council (SEC), which oversees Ministry and government-sponsored schools and provides broad education policy for the country. The SEC consists of three institutes: the Higher Education Institute (HEI), the Education Institute, and the Evaluation Institute. Figure 1.1 shows the organizational structure of the institutes and their relationship with the Ministry of Education from the inception of the reform in 2002 to 2006.[4]

Education Institute. The Education Institute oversees the development and opening of government-funded Independent schools through the Independent School Office. Unlike traditional Ministry schools, the Independent schools have the authority to operate under their own budgets and hire and train staff. We refer to the newly developed Independent schools, Ministry of Education schools, and private Arabic schools as *government-sponsored* schools. By the 2006–2007 academic year, there were 46 Independent schools in Qatar—approximately 20 percent of the total 227 publicly funded schools (Qatar Ministry of Education, 2007). Twelve Independent schools opened in 2004, and by 2008–2009, 85 Independent schools were in operation.[5] While the Independent schools maintain operational autonomy from the Ministry of Education, subject to the terms of the contract signed with the Education Institute,

Figure 1.1
Organizational Structure of the Education for a New Era Reform, 2002–2006



RAND *TR620-1.1*

[4] In May 2006, the Minister of Education, Sheikha Ahmed Al-Mahmood, was named as the Secretary General of the SEC, making her responsible for the operation of the Ministry of Education and the three Institutes. At this time, the Research Office was expanded to become the Office of Strategic Planning and Research and was placed directly under the auspices of the Secretary General's office.

[5] During the 2004–2005 school year, 9,107 students were enrolled in Independent schools, 68,287 in Ministry schools, and 62,507 in private Arabic schools (Qatar Ministry of Education, 2005).

the Ministry continues to directly operate the remaining traditional government schools (as shown in Figure 1.1). The Education Institute also developed curriculum standards for the core academic subjects of mathematics, science, Arabic, and English as a foreign language through the Curriculum Standards Office (CSO). Ministry and private Arabic schools do not follow the newly developed curriculum standards. Instead, they continue to follow the Ministry-developed curriculum.

Evaluation Institute. The Evaluation Institute provides information about government-sponsored schooling in Qatar, including Ministry, private Arabic, and Independent schools. Since 2004, it has developed a set of assessments based on the curriculum standards. Results from these assessments are fed into a national database, the Qatar National Educational Database System (QNEDS). The Evaluation Institute has also developed a set of surveys designed to capture contextual influences of test performance, which are another part of QNEDS. Each year, the questionnaires are administered to all students, their parents, teachers, and school administrative staff in the targeted schools in Qatar. The student assessment information can be linked to the student, household, teacher, social worker, and school administrator information to allow an array of research possibilities. The Evaluation Institute produces annual school report cards for each participating school, providing information from the surveys and assessments. The report cards were issued for the first time in April 2006, using data from the 2005 assessments and surveys. Within the Evaluation Institute, the Student Assessment Office (SAO) has responsibility for designing and developing the standards-based assessments and the School Evaluation Office (SEO) designs and implements the school-based surveys and is responsible for the development and dissemination of the school report cards. The Data Collection and Management Office (DCMO) administers the assessments and surveys and warehoused the data until 2007, at which point a separate entity under the assistant director for the Evaluation Institute took over the responsibility.

Higher Education Institute. HEI, established in March 2005, manages Qatar's postsecondary scholarship system as a complement to the other institutes, provides career guidance for Qatari students, and certifies private higher-education institutions wishing to operate in Qatar. As part of its remit, HEI administers scholarships and identifies top universities, degree programs, and short-term professional development courses in Qatar and around the world for HEI scholarship applicants. HEI also determines target specialties for scholarship recipients.[6]

Supporting Accountability Through the Student Assessment System

A key component of the ENE standards-based accountability system is the student assessment system. Before the reform effort, testing in Qatar had limited uses. The school-level tests described earlier did not facilitate systematic comparisons of schools' performance. Testing as a whole did not allow for the tracking of student growth over time, give any indication of Qataris' skills relative to those of students in other nations, or provide diagnostic feedback to teachers. In addition, the tests assessed factual, subject-matter knowledge rather than critical thinking, problem solving, and other more cognitively demanding skills, all of which the ENE reform initiative aims to promote. The testing system in Qatar was therefore seen as inadequate to forward the broad goals of ENE. To promote two key principles of the reform—holding

[6] Qatar's leadership has initiated other reforms to the country's postsecondary education system: reforming the one national university, Qatar University, and inviting U.S. universities to open branch campuses in the newly developed Education City.

schools accountable for the education of schoolchildren and encouraging parental choice—it was clear early on that Qatar needed an assessment system that would allow individual student–level longitudinal and cross-school and international comparisons so that the educational progress of the students and schools could be followed over time. The Qatar Student Assessment System (QSAS) was developed to promote these two principles and is based on the newly developed curriculum standards. To meet the demands of the standards, its design incorporated a number of assessment formats and types, depending on the skills or knowledge to be measured.

The accountability system underpinning the ENE reform rests on the notion that information generated by the QSAS would help improve student learning and increase stakeholder involvement and engagement in the Qatari education system by promoting interactions and conversations among students, parents, teachers, school administrators, business leaders, university admission officers, and policymakers. School-level results from the assessments would be made available on school report cards developed by the Evaluation Institute's SEO.[7] Various stakeholders would use the results of the QSAS to make informed decisions about the progress of students, schools, and the reform effort as a whole. Thus, the QSAS had to provide the following:

1. publicly available information about school performance to motivate school improvement and promote informed parental choice
2. feedback to teachers, helping them tailor instruction to support the needs of their student bodies
3. detailed information for policymakers about the education reform's progress in general and, specifically, about the Independent schools' performance for accountability purposes.

Purpose, Approach, and Limitations of This Report

From 2002 through 2005, RAND assisted Qatar with the implementation of the early stages of the reform effort. For three years, RAND project team members assisted the staff of the institutes and the SEC to build the institutes and design and develop the various components of the reform plan. As part of the implementation effort, RAND team members in the United States and Qatar worked closely with the SAO to design the QSAS.

This report provides a historical account of the early stages of the design and development of the QSAS and the development and administration of its core component, the Qatar Comprehensive Educational Assessment (QCEA). The QCEA is a standardized summative assessment with multiple-choice and open-ended questions, administered to students at the end of each academic year, and based on the Qatar curriculum standards. We provide a look at recent QCEA results and offer lessons learned from our experiences that may be of use to policymakers or test developers in other countries that are considering instituting a standards-based assessment system.

[7] Since 2006, the SAO has augmented the school report cards with reports for students and teachers that provide individual- and classroom-level results, respectively.

This report relies on three information sources. First, we reviewed early documentation written by RAND team members from July 2002 through February 2003, prior to the hiring of staff in the SAO. This documentation included reviews of the fields of accountability, standards-based education, and assessment theory and practitioners' guides to developing assessments. Second, we reviewed the minutes of all meetings held between July 2002 and July 2005 among RAND team members; Evaluation Institute leaders; SAO, DCMO, and CSO staff; and representatives from the contractors that assisted in the development and administration of the assessments. To fill in gaps in the minutes, we relied on our memories and checked with other meeting attendees if any discrepancies were found. Third, we reviewed internal memos—from both RAND and the SAO—to frame the decisionmaking processes of design features and policies of the QSAS.

Given the historical nature of this report, it is important to keep in mind several limitations. First, this report is limited in scope. Although it provides details on the nature of the testing operation and its properties, it is not meant to be a testing technical report. The test development companies have provided such technical reports directly to the SAO. Moreover, we did not assess the validity of the tests themselves. Although a valuable and necessary part of any testing effort, such an analysis is beyond this report's scope. Second, we offer our own perspective on the design and implementation process, which is not necessarily shared by others (e.g., other staff of the Evaluation and Education Institutes or the test contractors). A third limitation is that it was difficult at times to uncover who within the governance structure of the reform effort made certain decisions about the assessment system. It is not the purpose of this report to assign decisions to specific actors; therefore, if a specific decisionmaker is unknown, we note only the decision and when it was made.

Organization of This Report

In Chapter Two, we describe the design of the QSAS as originally envisioned. Chapter Three discusses the development of the structure and content of the QCEA in 2004, 2005, and 2006, focusing on key assessment development decisions. Chapter Four details the scoring and reporting of the QCEA in 2004, 2005, and 2006. This chapter provides the first look at results from the QCEA. We conclude in Chapter Five with a discussion of lessons learned and key recommendations for Qatar's Evaluation Institute as it continues to refine and implement the QSAS, as well as for other countries considering developing a standards-based assessment system at the national level.

The report also includes three appendixes. Appendix A tabulates the assessment elements considered in the development of the QSAS design. Appendix B presents a detailed discussion of the process of aligning the assessments to the curriculum standards. Appendix C provides 2005 and 2006 performance-level test results from the QCEA.

CHAPTER TWO

Design of the Qatar Student Assessment System: A Work in Progress

In August 2002, the SEC asked RAND to conduct background research on possible options for a standards-based assessment system in Qatar. It also asked for support to administer assessments in the spring of 2004, before Independent schools opened in September 2004. RAND organized a team (with specialties and experience in psychometrics, behavioral psychology, statistics, economics, and sociology) to work with the director and staff of the Evaluation Institute to design the QSAS. The director of the SAO and a small core group of staff were hired in 2003. At that point, the RAND team and SAO staff collaborated extensively to develop the QSAS and its core component, the QCEA—a standards-based, paper-and-pencil assessment battery. From 2003 to 2004, the SAO and RAND teams also worked to staff the SAO, build relationships with the other offices in the Evaluation Institute, design the QSAS, and hire the test development contractors that would develop the QCEA items. This chapter describes the process by which SAO staff, the RAND team, and members of the Evaluation Institute envisioned and designed the QSAS and the QCEA.

The QSAS Design as Initially Envisioned

Purpose and Uses of the QSAS

The SAO and RAND teams envisioned that the QSAS would support the ENE reform in a variety of ways. First, the assessment system should be flexible and responsive to the needs of the various stakeholders and potential changes in the curriculum standards that may occur over the years. Second, the results of the assessments should be clearly communicated to various stakeholders to enhance the transparency of the system. Third, test items should be fair and closely linked to the curriculum standards to ensure that the assessments accurately and objectively assess student learning in key content areas, problem-solving skills, and critical-thinking skills using multiple measures. Fourth, the test results should be valid and reliable. In turn, the results of the assessments should enable stakeholders to evaluate student readiness for local and international higher education and the workforce, as well as assess student achievement in relation to local, national, and global needs and developments.

Through a series of meetings in 2003, the SAO and RAND teams decided on the purposes and uses of the QSAS. As discussed in Chapter One, the goal was for a variety of educational stakeholders to use the results of the QSAS to make informed decisions about the progress of students and schools, which would provide the following:

1. publicly available information about school performance to motivate school improvements and promote informed parental choice
2. feedback to teachers, helping them tailor instruction to support the needs of their student bodies
3. detailed information for policymakers about the education reform's progress in general and, specifically, about the Independent schools' performance for accountability purposes.

Format and Composition of the QSAS

Once the purposes and uses of the QSAS were determined, the SAO and RAND teams discussed and debated a variety of assessment types and formats to determine those that would best fit the proposed purposes and uses. It was clear from the onset that no single assessment would suffice. The SAO decided that the QSAS would consist of multiple components to provide a comprehensive picture of students' abilities in a standardized manner. By using a variety of assessment components, the QSAS could test student knowledge of the standards in an appropriate format and method.

We decided that the framework should include both summative and formative assessments. Summative assessments are tests that gather information about learning after the learning has occurred. In the context of a classroom, summative tests are typically used for the purpose of assigning grades to students. Formative assessments gather information about learning as learning is taking place during a school year. In a classroom, formative assessments are typically given periodically to allow teachers to monitor students' progress toward learning the curriculum standards and to adjust instruction accordingly. Some national systems employ a combination of standardized formative and summative assessments developed both by teachers and by a professional testing company (Linn, 1998; Stapleman, 2000).[1] Through national, formative, and summative standardized assessments, school administrators and policymakers could analyze school-level achievement trends through the years. An added benefit would be that, by training teachers to develop formative or summative assessments, they would become more familiar with the standards and, in turn, be better equipped to teach them.

So that Qatar's education policymakers could gauge Qatar's students' knowledge, skills, and progress relative to those in other countries, the SAO also considered Qatar's participation in international assessments. In 2006, Qatar participated in the Programme for International Student Assessment (PISA) and the Progress in International Reading Literacy Study (PIRLS). In 2007, Qatar participated in the Trends in International Mathematics and Science Study (TIMSS).[2]

---

[1] In the UK, for example, students are tested at the end of key stages, which correspond to age levels. Thus, at the end of key stage 2 (age 11), pupils are assessed by their teachers and by standardized national tests in four core subjects (English, math, science, and information technology). Parents receive both sets of scores.

[2] The testing programs differ with respect to scope and content assessed and the age or grade level they target. PISA is a battery of tests that requires 15-year-old students to apply their science, mathematics, and reading skills to "real-life" situations. PIRLS is administered in the fourth grade and assesses a broad range of reading literacy skills, including the ability to retrieve specific information, make inferences, interpret ideas, or evaluate the text. TIMSS, which is administered in the fourth and eighth grades, reflects general curricula likely to be found in most schools and provides information about how well students have learned the mathematics and science concepts to which they have been exposed in school. Each assessment tests students of the pertinent ages or grades in all government-sponsored schools of the corresponding levels in Qatar.

We next turned our attention to the specific components of the QSAS. The new Qatar curriculum standards were due to be completed in January 2005. The SAO therefore could not finalize the QSAS design until after that time. However, the team was able to review draft standards and discuss them with the test development contractors. The solution that we adopted was to weigh the benefits and disadvantages of a number of assessment methods, testing strategies, and item formats that could be modified and adjusted once the standards were finalized. Because the standards were developed to emphasize critical thinking and problem solving, the assessment system needed to measure these skills in addition to subject-matter content. Appendix A lists the assessment components that were considered and their pros and cons.

Item Format. Because constructed-response items and other open-ended formats take more time to complete, tests composed solely of these types of items cannot cover as large a range of skills and knowledge as multiple-choice items in the same amount of time. However, the efficiency of the multiple-choice format is offset by its limitations in measurement of certain higher-order skills, such as writing proficiency. Moreover, multiple-choice items encourage the use of less desirable solution strategies (e.g., excluding wrong alternatives instead of knowing the right answer, guessing). For these reasons, other assessment methods, such as portfolios, constructed-response items, and performance-based measures, were also considered. In addition, these other methods are better suited to assessing critical thinking and problem solving (Bennett and Ward, 1993).

Assessment Method. One assessment method considered was performance-based assessments, which require students to perform a task (e.g., a scientific experiment) or to generate an extended response (e.g., a research paper). Another option considered was a portfolio, which is a collection of a student's work that typically shows his or her progress through a school year or term and includes his or her thoughts and reflections on the work. Often, a panel of teachers uses a standard set of criteria to grade the portfolio.

Testing Strategy. We also considered the delivery of assessments to students. We weighed the benefits and disadvantages of paper-and-pencil level tests, which are tests that consist of different forms (e.g., low-, medium-, and high-proficiency-level forms) with content that is more closely matched to an individual's proficiency level. Ideally, the proficiency-level forms overlap, sharing a common measurement range and some test items. Each deals with the same concepts and topics but at differing levels of complexity. We also considered computer-adaptive testing (CAT), in which questions are administered to the examinee according to his or her demonstrated proficiency. Based on answers to earlier items, the test presents either harder or easier test questions that better fit the proficiency level of the examinee.

The one exception is PISA, which was administered to 15-year-old students in all schools in Qatar—both government-sponsored and private schools.

Apart from the international comparison of student performance, the participation in the three major international comparative studies in 2006 and 2007 allows Qatar to establish a baseline against which it can track the country's performance outside the QCEA in the corresponding subject areas across study cycles. (For example, in PIRLS, students are given 80 minutes to read two long text passages and answer questions. The specific focus and scope of this particular task makes it difficult to accomplish through annual national tests.)

QSAS and QCEA Development Issues: Turning Design into Practice

The process of developing standardized, standards-based assessments was an enormous task that had never before been undertaken in Qatar—or in any country in the region. A number of development issues had to be addressed early on. In this section, we itemize the key issues with which the SAO and RAND teams grappled, noting decisions, their rationale, and how Qatar's educational policymakers readdressed these issues through the years as the QSAS and QCEA were implemented.

Where to Start?

The first batch of Independent schools opened in September 2004. Educational leadership in Qatar expected to have a standards-based assessment system in place by the end of the 2004–2005 academic year so that students' learning vis-à-vis the newly developed curriculum standards could be appropriately measured. By 2005, the SAO had laid out the QSAS's goals, purposes, uses, and design features, but given the tight timeline, a fully developed assessment system could not be put into place. We therefore focused our efforts on the planning and development of the QSAS's core component—the QCEA—a summative, paper-and-pencil assessment administered at the end of the school year.

Development and implementation of the QCEA was prioritized over other proposed components of the QSAS for three reasons. First, the QCEA was to be the largest and most comprehensive component of the QSAS. Other summative assessments of the QSAS were to measure skills and knowledge not adequately assessed in the QCEA. Moreover, the plan was for the QSAS's formative assessments to be administered throughout the school year, depending on a school's needs or delivery of the standards. Given the size and scope of the QCEA, coupled with the fact that the SAO would need to coordinate the delivery of formative assessments to Independent schools that had not yet opened, it made sense to start with the QCEA and then, over time, develop the other components. Second, in 2003 and 2004, the SAO worked with a skeleton crew of three staff members—a director, a psychometrician, and an evaluation specialist seconded from the Ministry of Education—which limited the office's capacity to focus on the implementation of the QCEA alongside the implementation of other components of the QSAS. Third, the SAO and RAND teams and the test developers worked with draft standards until they were finalized in 2005. Final decisions on the overall design of the QSAS therefore could not occur until the standards were finalized.

Which Students Would Be Part of the QSAS?

A central issue for Qatar's education policymakers was whether to incorporate students in the QSAS who may not have been exposed to the new curriculum standards. Independent schools were required to follow the newly developed curriculum standards, yet students in the Ministry and private Arabic schools continued to rely on the Ministry of Education curriculum. The concern was that the assessments would not be directly linked to the curriculum taught in classrooms; teachers, students, and parents might think that the tests were therefore unfair. Administrators of Independent schools, who were free to choose any curriculum for their schools as long as it followed the standards, might also perceive the assessments as unfair because they were not directly linked to the classroom curriculum.

The SEC decided that the assessments would be administered to all students in government-sponsored schools: Independent, Ministry, and private Arabic. The rationale for

this decision was that the assessments would be aligned with the new curriculum standards—and not with classroom curricula. Test items' difficulty level or depth of knowledge would therefore not be matched to students' present learning capabilities or to a specific curriculum, but to the skills and competencies as written in the standards.

This decision was based on two reasons. First, as mentioned in Chapter One, the *choice* principle of the reform called for parents to make informed choices about where to send their children to school. Parents are able to retrieve information about government-sponsored schools in Qatar through the school report cards that the Evaluation Institute publicly disseminates. As part of the reform design, results from different components of the QSAS would feed into the report cards.[3] Therefore, it was vital to the success of the reform that the standards allow parents to have access to data on all students' and schools' progress. It was also important for education policymakers and researchers to have information on all students' learning of the standards, not just a subset of the student population, to measure the progress of reform—that is, whether students in the country were learning more over time. Second, the curriculum standards are meant to be the national standards of Qatar, against which all students' knowledge and skills are to be measured. Students should therefore be exposed to skills, knowledge, and tasks embodied in the standards. Decisionmakers hoped that inclusion of all students in the QSAS would propel the Ministry of Education and administrators of private Arabic schools to eventually adopt the curriculum standards or at least promote critical thinking and problem solving within the Ministry's curriculum, even if this meant that the testing operation would, in some sense, drive curriculum.

However, in November 2006, after three administrations of the QCEA and before other components of the QSAS could be fully implemented, the SEC decided to limit Evaluation Institute testing to only those students who were enrolled in the Independent schools, starting with the 2007 application. Only Independent schools, therefore, would have results from the assessments on their school report cards. Information from surveys administered to students, their parents, and school administrators in all government-sponsored schools would continue to be placed on school report cards for Independent, Ministry, and private Arabic schools.

This decision was made for a number of reasons. First, the SEC was deeply concerned about unfairness to students in the Ministry and private Arabic schools, whose curricula were not based on the Qatar curriculum standards but on the curriculum developed by the Ministry of Education. Central to decisionmakers' concerns was that the Ministry and private Arabic school students were being tested on content and competencies to which they may not have been exposed in their class work. This lack of opportunity to learn may, in turn, have had effects on the validity of the QCEA. Second, the SEC decided that all Ministry schools would eventually convert to Independent schools, although the timeline for conversions had not yet been fixed at the time of the decision.[4] Therefore, the SEC intended for all students in Ministry schools to eventually participate in the Evaluation Institute assessments—once Ministry schools converted to Independent schools.

The decision to include only Independent school students has a number of benefits. First, it slightly lowers the cost of administering the QCEA. However, the cost difference is

[3] The report cards also have information taken from a set of school administrator, student, teacher, and parent surveys developed by the SEO within the Evaluation Institute.

[4] In March 2007, the director of the Education Institute announced that all Ministry schools would become Independent schools by 2011.

negligible, considering that DCMO staff still need to administer surveys to students in all government-sponsored schools. Second, it lowers the stress and burden on students in Ministry and private Arabic schools. Finally, it improves the public's perception of tests.

Nevertheless, a number of short-term and long-term drawbacks exist that may not have been considered when the SEC made this decision. The Qatar curriculum standards are considered the national standards against which the progress of all students in government-sponsored schools should be measured. Having one national standard provides a number of benefits to measuring the progress of the reform effort and of students over time. By limiting testing to only a subpopulation of students in government-sponsored schools, the reform effort itself will be limited because education policymakers will be unable to (1) compare students in government-sponsored schools over time, (2) measure the progress of the reform effort over time, (3) compare government-sponsored schools to each other, and (4) understand how much an effect Independent schools have on student learning for those students who move from a Ministry or private Arabic school to an Independent school. QCEA scores provided a baseline for researchers and policymakers to know how much students improve once they move to an Independent school.

What Would Be the Structure of the QCEA?

The third issue was how to structure the QCEA. In order to assess the effects of the reform on student achievement, RAND and the Evaluation Institute felt that it was important to have a baseline of student performance before the reform was implemented (i.e., test them at the end of the 2003–2004 school year). However, as mentioned previously, the curriculum standards would not be finalized until January 2005. Given this constraint, the SAO and RAND teams determined that the QCEA would need a two-stage design. The first stage would provide a "snapshot" of student's general knowledge relative to general standards determined by test development companies contracted by the Evaluation Institute in mathematics, science, and English as a foreign language and, for the first time in the region, a national standardized test in Arabic language. In the second stage, the assessments would be aligned to the new curriculum standards once they were completed.

Given the abbreviated timeline to develop the QCEA in 2004, as mentioned earlier, the SAO and RAND teams decided that the test would consist of multiple-choice questions and some essays: Students in grades 7–12 write essays in Arabic, while students in grades 10–12 write essays in English. In 2005, the format was expanded to include constructed-response questions. The process to align the QCEA with the standards started with the 2005 administration.

How Would QCEA Results Be Used?

The fourth issue was how results from the QCEA would be used. As part of the reform effort's accountability mechanism, schools are held accountable by parents' decisions to enroll their children in schools of their choice. As noted in Chapter One, results from the assessments, along with other information about schools, could help inform parents' decisions about which school to send their child to. Furthermore, Independent schools could use school-level assessment results to determine where students may need extra support or assistance. The SAO advised participating schools not to use results from the assessments as the sole determinant for student promotion or teacher rewards because it is generally recognized that exams cannot adequately capture the full range of capabilities deemed important by stakeholders. Concerned

that Independent schools may select the best students from among entering Ministry of Education students, the Education Institute stipulated an admission policy based on a student's nationality and parents' employer. Independent schools are not allowed to use QCEA test results for admission purposes.[5]

In Which Language(s) Would the QCEA Be Administered?

The fifth issue was the language in which the QCEA would be administered. In 2004, 12 Independent schools opened. At the time of opening, the Education Institute allowed them to select any language of instruction as long as the curriculum followed the standards. Five schools opted to teach their students in English. Stakeholders raised the question of whether the QCEA should have both an English and an Arabic version to accommodate the schools that taught in English. The main argument against assessments being conducted in a language other than the language of instruction is the difficulty in transferring content terminology, even when the language of assessment is the student's stronger language (Abedi, 2004). This would apply, for example, to science terminology (e.g., students learn the term *photosynthesis* only in English) or the ease with which one carries out mathematical operations (e.g., the "Arabic" numeral *6* versus the "Hindu-Arabic" numeral ٦, which is used in the Gulf region). Given the difficulty and time required to adequately develop parallel versions of an assessment in another language (including the research needed to establish the statistical and linguistic equivalence of those forms) and the relatively small numbers of students taught in English, the SAO opted to develop the QCEA solely in Arabic for the first years of the testing initiative.

Even with this decision, there remained a concern about whether students taught in English would be at a disadvantage when taking the assessments in Arabic. In October 2004, the SAO commissioned a series of papers by leading scholars in the field of language in teaching and testing. Among other topics, the issue of interpretability of scores obtained by students whose instructional language differs from the language in which they are tested was discussed. In these papers, the authors raised the issues of equity and validity. To evaluate whether students' performance in mathematics and science tests are unduly hindered because they are tested in Arabic instead of English, as part of the 2005 QCEA, the SAO carried out an additional assessment of fifth graders in schools that self-identified as adopting English as the primary language of instruction for mathematics and science. With the support of CTB/McGraw-Hill (CTB), RAND produced a design for the field-tested portion of the 2005 QCEA mathematics and science exams in which each student in the English-instruction schools received one of four test forms. Students in the schools of interest were given additional fifth-grade field-test items that were not part of the individual QCEA form, administered in English and Arabic or Arabic with an English glossary. Combining these additional items with their responses to the main portion of the QCEA yielded a cross-classified design of students-to-item-language combinations that allowed for the identification of the effect of the language of administration on item difficulty. The SAO carried out an internal study of the data produced by this additional assessment. Results suggest that students instructed in English did not encounter any disadvantage when tested in Arabic in mathematics, but that in science, students experienced a small disadvantage when tested in English rather than their native Arabic (Broer, Froemel, and Schwarz, 2007).

---

[5] See Qatar Supreme Education Council (undated[a]), for details on the admission policy for Independent schools.

In the fall of 2005, the SEC informed the Independent schools that they needed to gradually move toward teaching mathematics and science in English. The Independent schools expressed concerns that the language differences between instruction and testing would place their students at a disadvantage. Although the SAO's results from the experimental administration conducted in grade 5 during the 2005 QCEA suggest that the language of administration did not have an impact on mathematics scores and that students who were taught science in English would likely have obtained a lower score had the QCEA been conducted in English,[6] the Evaluation Institute decided that the 2006 QCEA would be offered in both English and Arabic, depending on the school's language of instruction. Schools had to specify the language in which they were teaching mathematics and science, and their students received test forms that corresponded to the language of instruction. Seventeen schools selected the English version of the 2006 mathematics and science QCEA, resulting in 4,610 student scores for mathematics and 4,438 student scores for science.[7]

What Would Be the Delivery Method of the QCEA?

RAND initially envisioned the QCEA as a computer-adaptive test, in which items are administered according to each examinee's demonstrated proficiency. Based on answers to earlier items, the test branches to progressively harder or easier test questions. By presenting items of appropriate difficulty, a computer-adaptive test provides more precise estimates of an examinee's proficiency than do more traditional paper-and-pencil tests, and it does so in a shorter amount of testing time.

However, to put CAT into practice, schools must have sufficient capacity with regard to computer availability and quality of hardware and software. Qatar, like most countries, currently lacks the infrastructure to support CAT on a nationwide basis. Other issues also came to light that would prohibit the implementation of CAT in Qatar in the early years of the testing program:

- *Item-pool development:* Even if all logistic concerns had been solved, to properly implement CAT to its full potential, large item pools are needed. In the early years of the reform, this would not have been feasible.
- *Item parameters:* CAT requires that item parameters, particularly the item-difficulty parameter, be fairly stable—something that cannot be expected in the early years of the reform initiative when most items are continuously pilot tested.
- *Item types:* CAT works very well with multiple-choice items or one-point constructed-response items in which the solution is entered. For example, in mathematics, the student enters the solution directly. However, CAT cannot work with open-ended items that require scoring. A large portion of the items used in the QCEA require scoring, since those items are designed to capture the cognitively complex processes demanded by the standards. This means that if one were to introduce CAT on those open-ended items,

[6] These results may, however, hold only for the transition period during which students are adapting from an Arabic- to an English-taught curriculum.

[7] Because of the speed with which the English forms had to be prepared, there was no time for a pilot test. To ensure consistency and comparable difficulty, the English forms used the originally developed items, which were then reviewed again against the Arabic versions that had previously been used on the QCEA.

the depth-of-knowledge alignment may suffer. Therefore, CAT can be used only with multiple-choice items or limited constructed-response items.

In light of the lack of capacity to support CAT on a wide-scale basis and the other issues noted here, the QCEA is currently paper-delivered and administered in classrooms with a proctor.

Which Grades Would Be Tested by the QCEA?

The seventh issue concerned the grades that would be tested. With support from RAND analysis, the Evaluation Institute decided that all eligible students in all grades (1–12) would be tested with the QCEA (except those who were learning disabled or had other special needs). This decision was made for three reasons. First, it reflected the SEC's priority to have longitudinal information at the individual student level. Having data on students in all grades in the early years of the reform allows for tracking of students' progress. Second, the number of students in Qatar's government-sponsored schools is relatively small (approximately 89,000 in 2004). To provide school-level information by grade on school report cards and to use appropriate modeling techniques to gauge school-level improvements, it would be best to have data for all the students in a school, rather than for a sample. Third, student-level test results would help students, parents, and teachers assess the strengths and weaknesses of individual students.

After the administration of the 2005 QCEA, however, the SAO raised questions about testing students in all grades and decided to stop testing students in grades 1, 2, and 3 altogether. There were three reasons for this decision. First, test administration in 2005 for students in grades 1–3 may not have been standardized because proctors read the test questions in all four subjects to compensate for the fact that students' reading skills were not well developed in those grades. Proctors varied in their spoken modern standard Arabic or English proficiency, which may have confused students. Second, the reading of test questions may have differentially benefited students, depending on their listening skills. This affected the reliability of the test. The test contractors found that test reliability for these grades was well below acceptable levels due to poor construct validity. That is, it was impossible to differentiate which cognitive skill the test was measuring: reading comprehension, oral comprehension, subject knowledge, or a combination of the three. Third, students' lack of familiarity with standardized testing in general and anecdotal evidence from test observations about some students experiencing test anxiety were further reasons to doubt the usefulness of testing in grades 1–3. Because of these issues, guaranteeing the test's standardization for students in these grades was deemed too difficult.

The SAO also decided to eliminate the QCEA for grade 12. In 2004 and 2005, the grade 12 QCEA suffered from very high absentee rates (2005 absentee rates for mathematics were 23 percent; for science, 32 percent; for Arabic, 24 percent; and for English, 40 percent). It was felt that the students at this grade level were more concerned with their impending exit exams than with the QCEA.[8] At the time, the exit exam scores were the most important assessment factor for students to compete for scholarships and university entrance.

[8] Ministry and private Arabic schools continue to administer an exit exam to 12th-grade students. Starting in June 2007, the Evaluation Institute certified the graduation of 12th-grade students in Independent schools through a set of subject tests based on the QCEA and school-level accomplishments.

Table 2.1 summarizes the changes in the design features of the QSAS and QCEA from 2004 through 2007.

Table 2.1
QSAS and QCEA Design Changes, 2004–2007

Assessment Feature	2004	2005	2006	2007
QSAS				
Students participating	Ministry Private Arabic	Ministry Private Arabic Independent	Ministry Private Arabic Independent	Independent
Standards tested	General standards determined by test development contractors	New Qatar curriculum standards	Same as 2005	Same as 2005
Other standardized assessments administered	None	PISA and PIRLS pilot tested	PISA and PIRLS; TIMSS pilot tested	TIMSS
QCEA				
QCEA general purpose	"Snapshot": preliminary information about the proficiency distribution of students in Qatar according to international standards	First consolidated baseline for new standards-aligned tests	Baseline + 1 year	Baseline + 2 years
Testing language for mathematics and science	Arabic	Arabic, with language of assessment study for fifth graders in five Independent schools	Arabic and English available depending on school's choice	Same as 2006
Delivery method	Paper-and-pencil/ multiple-choice + some essays	Paper-and-pencil/ multiple-choice + constructed-response + essays	Same as 2005, with more than half being constructed-response items	Same as 2006
Grades tested	All subjects: 1–12	Arabic, English, mathematics: 1–12; Science: 4–12	All subjects: 4–11	Same as 2006

CHAPTER THREE

Implementing the QCEA in 2004, 2005, and 2006: Test Development and Administration

Decisions about which skills the QCEA would assess were determined jointly among the SAO and RAND teams, the test development companies contracted by the Evaluation Institute (CTB and Educational Testing Service, or ETS), and—for the 2005 QCEA—the organization that had contracted with the Education Institute to create the curriculum standards, the Centre for British Teachers (CfBT, now the CfBT Education Trust). Around the time that the test developers were contracted, the Evaluation Institute hired the National Opinion Research Center (NORC) to facilitate the printing and administration of the tests and the surveys fielded to students, parents, teachers, and school administrators. A number of measures were put into place to audit the alignment of the QCEA with the standards for 2005 and beyond. This chapter documents the development of the QCEA in the first years of the reform, covering January 2003 through April 2006; the alignment process of the 2005 QCEA with the Qatar curriculum standards; and the administration of the 2004 and 2005 QCEAs. It also discusses changes in the 2007 implementation as planned in April 2006.

2004 QCEA: First Year of Standardized Testing

The process of developing the QCEA for April and May 2004 administration began with the recruitment of test developers in January 2003. RAND and the Evaluation Institute sent a request for proposals to a selected group of four testing companies with strong qualifications, including a record of developing assessments on a large-scale basis, extensive knowledge of computer-delivered and/or computer-adaptive testing, and experience with test development in an international context. Each company was invited to the RAND offices in Santa Monica, California, to learn more about the reform efforts and the envisioned assessment system. After extensive review of their proposals to develop assessments in all four subjects—two companies, ETS and CTB—were invited for further meetings in Doha, Qatar, in June 2003. At those meetings, based on the recommendations forwarded by the RAND team, the Evaluation Institute chose ETS to develop the tests of Arabic and English as a foreign language and CTB to develop the mathematics and science tests. The companies were contracted to develop the first-year QCEA snapshot, to be administered in April and May 2004, as well as to align their tests with the new curriculum standards once they were completed.[1]

[1] Although the request for proposals only called for the development of the first-stage "snapshot" test for the 2004 QCEA, the test development companies argued that the testing operation would have a smoother and faster transition from the first stage to the second stage of aligned tests if they could work on the alignment immediately after the administration of the 2004 QCEA. Aligned assessments could therefore become available starting in 2005.

The test developers had approximately eight months to develop, pilot, and finalize items from the time they were hired in June 2003 to January 2004, when the camera-ready copies were due to the Evaluation Institute. To accommodate the short time frame for development, the fact that the mathematics and science items had to be produced in Arabic, and the fact that the 2004 QCEA would provide a simple snapshot of student performance across a broad spectrum of skills and content areas, the SAO and RAND teams decided that the test format would consist of multiple-choice questions and some essays. The English assessments included a single essay question for students in grades 10–12. For the Arabic assessments, an essay question was administered to students in grades 7–9 and a different one to students in grades 10–12. The understanding was that this format would expand to accommodate constructed-response questions or other item formats when the QCEA became aligned to the standards.

Initially, the test companies were asked to create a snapshot test with preliminary versions of the curriculum standards but that nonetheless could enable comparison of student performance in 2004 with performance in subsequent years. However, in October 2004, when the SAO, RAND, ETS, CTB, and CfBT examined the overlap of the 2004 individual QCEA items with the draft curriculum standards, it became evident that the overlap was insufficient to justify a valid comparison between 2004 results and the subsequent standards-aligned QCEA tests. Thus, the SAO decided that future QCEAs would not link back to the 2004 QCEA, but rather to the 2005 QCEA.[2]

Item Development

Given that the Qatar curriculum standards were not yet complete, the Evaluation Institute requested that the test contractors develop tests that would broadly measure international standards of content knowledge and competencies. For the 2004 QCEA, the test developers therefore created or selected questions for which an "averagely able" population would achieve a mean proportion of 50 percent correct. ETS constructed new items for Arabic and English using item-writing experts in the United States and Jordan, where ETS had a relationship with Arabic language teachers to serve as item-writers. CTB linguistically and culturally adapted mathematics and science items from one of its existing large-scale programs (TerraNova) to Arabic.

To ensure the cultural appropriateness of the items, both ETS and CTB worked with education professionals from the Middle East region. In addition, Evaluation Institute staff members reviewed all items. Names that were not considered "Qatari" were changed, and all names on the English tests were transliterated from Qatari names so that students would know how to spell the names in English. Moreover, any situations that seemed culturally inappropriate were also changed. For example, one reading passage in the language tests included a situation in which a young man travels through the woods and comes upon a young girl alone in her house. They have a conversation, about which students were to answer questions. For Qataris, a girl alone with a strange young man would be considered culturally inappropriate and the situation was therefore changed. Another item featured illustrations in which girls wore skirts shorter than what would be considered appropriate in Qatar. These illustrations were revised to reflect Qatari societal norms. In addition, steps were taken to ensure that items were not offen-

[2] Since the 2005 QCEA required long testing times already, the SAO rejected the option of including additional non-aligned items just for the purpose of linking back to 2004.

sive or stereotypical. For example, many questions with only camels or falcons—two animals that are common in the country—were changed so that students could be exposed to a variety of animals from the region and—in older grades—from around the world.

After the cultural sensitivity review, the items were checked for clarity, grade-level appropriateness, and technical quality.[3] This was achieved through a usability test of the adapted TerraNova mathematics and science items and a pilot test of the newly developed Arabic and English items. CTB uses the term *usability* to describe the effort to gather information from feedback sessions and evaluate basic item quality measures. *Pilot* refers to the effort to gain sufficient information on item functioning to assemble future main application test forms. ETS was able to use the item statistics to select items for the spring main administration. CTB's data on usability of items allowed it to understand student capacities and how well students understood how to take a test in the multiple-choice format.

The usability and pilot tests were administered in October 2003 to grades 2–12. As it was early in the school year, students in each grade were given items that would appear on the end-of-year exam for the grade below; for example, second-grade students were given first-grade items. Because of the students' limited reading proficiency at the early elementary grade levels, proctors read aloud the first-, second-, and third-grade test items. Because it was necessary to include common items on test forms from adjacent grades, items that appeared on both the third- and fourth-grade tests were also read aloud to fourth-grade students. However, for the majority of test questions, fourth-grade students were required to read the items by themselves.

A total of 45 schools participated in the usability and pilot testing,[4] with an average of approximately 335 students per grade participating in each subject. To learn students' impressions of the items, CTB held focus groups with a selection of students who had just taken the tests. Administration and testing time was limited to one class period (45 minutes). Representatives from each test company, the Evaluation Institute, and the RAND team observed a number of the pilot and usability tests; in total, 24 observers attended at least one day of the testing. At the end of each testing day, NORC led a debriefing session with the proctors and observers, gaining valuable logistical information for the QCEA main administration.

A meeting was held in Doha in December 2003 at which NORC, ETS, and CTB briefed the Evaluation Institute on the findings from the usability and pilot tests. NORC found that the administration time allotted for the tests was too short. In some mathematics classes, administration procedures took up all but 10 minutes of the class period, leaving only a few minutes for the students to take the tests. NORC suggested that, rather than restricting the tests to one class period, the tests should be administered over two class periods. ETS and CTB reported on absentee rates, missing data rates, item statistics on the English and Arabic tests, and overall impressions of the value of the pilot and usability testing administration. Any items selected or developed for the main administration that were not in the pilot or usability test were submitted to the Evaluation Institute for cultural review.

---

[3] The SAO hired content experts in 2005 to check the fidelity of construct-specific item content. For the 2004 QCEA, the SAO relied on the test developers for this check.

[4] In general, each contractor was assigned three Ministry of Education schools of each gender at each grade level, with the students in each school taking tests in two subjects on successive days. Two private Arabic schools (one preparatory and one secondary) were added for each contractor. The remaining five schools were male primary schools, or model schools, which go through either fourth or fifth grade.

Developing the QCEA in 2005

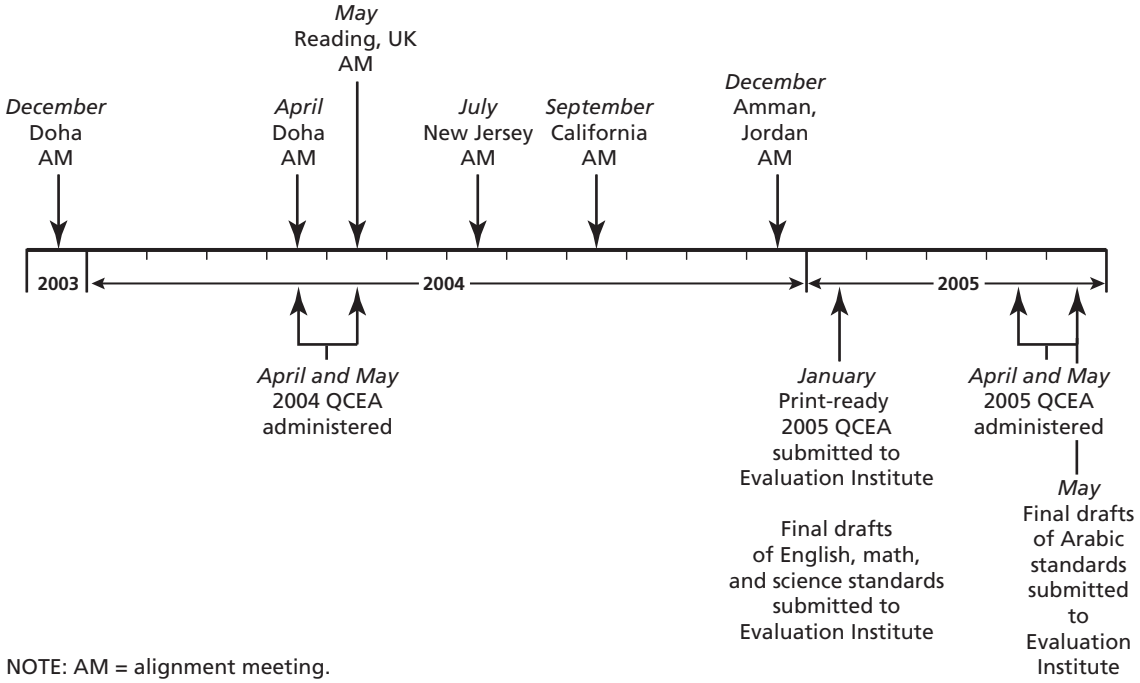
Aligning the QCEA to the New Qatar Curriculum Standards

One purpose of the 2005 QCEA was to provide baseline information about students' performance with respect to the curriculum standards. A major challenge for the test developers was to design items that were aligned to the new standards while the standards were still being developed. The deadline for final versions of the standards was January 2005, yet the test developers had to finalize the test items in November of 2004 and provide camera-ready copies of the tests in January 2005. Figure 3.1 provides a timeline of the alignment between the 2005 QCEA and the Qatar curriculum standards, showing the many alignment meetings in 2003 and 2004 and the relationship between those meetings, the drafts of the standards, and the administered QCEAs in 2004 and 2005.

The appropriateness of using the QCEA results to assess student performance relative to the curriculum standards depends in part on the quality and breadth of the alignment between the test items and Qatar's new curriculum standards. Both the content to be learned and the level of performance demanded by the test items must correspond to those established by the standards. Efforts were made to align the items with the draft standards, with particular attention paid to the "key" standards demarcated within the standards as most important for teaching.

During the item-writing phase, ETS and CTB met several times with CfBT, members of the Education Institute's CSO and the SAO and RAND teams to reach a shared understanding of the meaning of the standards. CfBT provided input and feedback on the items that

Figure 3.1
Timeline for Alignment of 2005 QCEA with Qatar Curriculum Standards, 2003–2005



NOTE: AM = alignment meeting.

RAND *TR620-3.1*

ETS and CTB had been developing, noting which standards the items appeared to measure and whether those items partially or fully measured the standards to which they had been mapped. ETS and CTB pointed toward ambiguity in some descriptions and problems in the progression of the standards from grade to grade. In developing the QCEA, CTB and ETS outlined a set of test specifications, or blueprints, that delineated the number of items needed to sample the standards. These test specifications were then used to show how items or tasks in the test matched the standards documents. The test specifications indicated that many of the standards were included on the QCEA. To further ensure the alignment between the 2005 QCEA and the standards, an alignment audit was commissioned by the SAO after the tests had been completed. Specific details of the alignment process between the 2005 QCEA and Qatar's new curriculum standards are presented in Appendix B. The alignment audit found that, although all items on the 2005 QCEA matched the standards, the test as a whole did not provide adequate coverage of the breadth and depth of the skills and content expected by the standards. This finding was to be expected, given that the design of the QSAS called for multiple assessment types to adequately cover the curriculum standards; the paper-and-pencil format of the QCEA could not cover all the standards.

One implication of the standards for test development is that 10th- through 12th-grade standards specify two levels, foundation and advanced.[5] In the first years of the QCEA, only the foundation level was tested. In these initial years of the reform, few if any of the Independent schools are expected to offer curriculum aligned to the advanced standards. Furthermore, few students will have had the prerequisite knowledge to pursue the more advanced standards, and the dearth of students pursuing the more challenging standards renders it difficult to obtain accurate estimates of proficiency. The SAO therefore decided that test developers should focus their efforts on aligning the QCEA with the foundation standards until more schools offered advanced coursework.

Changing the Format of the QCEA

The QCEA underwent a series of changes between 2004 and 2005 in an effort to align test content with the curriculum standards. These changes included the following:

- New item formats were added to each subject to assess a variety of skills that had not been assessed during the 2004 administration, and, as a result, more items were constructed-response. For example, half the questions in mathematics and science on the 2005 QCEA were short, constructed-response items, compared with none in 2004. Also, in Arabic

[5] Advanced standards allow students to acquire significant knowledge about specific subjects. Advanced students study the grade 10–12 foundation standards in grades 10 and 11 before moving to more advanced work in grade 12. The extra content at the advanced level focuses on new topics beyond the scope of the foundation level and provides more in-depth study of foundation-level material (e.g., harder problems, more demanding critiques of texts). According to an internal memo from CfBT, for mathematics and science, the end point for grade 12 foundation standards equates to the General Certificate of Education (GCE) advanced subsidiary (AS-level), or Scottish Highers, and the end point for grade 12 advanced standards equates to GCE A-level, or IB. For English as a foreign language, foundation standards are equivalent to the International English Language Testing System (IELTS) intermediate level 4.5 or a score of 450 on the Test of English as a Foreign Language (TOEFL). The English advanced standards are equivalent to the IELTS upper-intermediate level 5 or a score of 500 on the TOEFL. (The latter is equivalent to a modern foreign language, such as French GCE A-level for English speakers.) There is no real equivalent for Arabic as a mother-tongue language, but some equating can be done by looking at expectations for English proficiency among GCE A-level mother-tongue English speakers or French for IB mother-tongue French speakers.

and English, constructed-response items requiring short written responses were added to the tests for grades 4–12, in addition to the existing long essay-type constructed-response items in higher grades.[6]

- In mathematics and science, CTB moved away from adapting existing items in the TerraNova and developed new items specific to the curriculum standards. In addition to adding constructed-response questions in both subjects, CTB included items that assessed students' ability to use tools, such as calculators, rulers, and protractors, in mathematics.
- At the recommendation of CfBT, CTB conducted additional pilot testing within the main administration of the QCEA at the sixth-grade level to measure students' "mental mathematics" proficiency. Students listened to mathematics items played on a CD and had 5, 10, or 15 seconds to solve the problem without using paper and a pencil or a calculator. Most mental mathematics items assessed basic skills with fractions, operations, or measurements, but there were also some problem-solving exercises.
- Because the science standards for grades 1–3 are not based on specific content but instead focus on inquiry skills that are not easily tested via paper-and-pencil measures, students in those grades were not assessed in science on the 2005 QCEA. In Arabic and English as a foreign language, tasks requiring students to listen and answer questions were administered to students in all grades. Compared with the tests in the previous year, the 2005 measures give more emphasis to grammar in both languages.

Although the 2005 QCEA was expanded to include other formats, including constructed-response items as well as some performance-assessment tasks (in the form of listening tasks), it did not include portfolios or other types of performance assessments (e.g., those that assess oral communication or scientific inquiry through hands-on scientific experiments) because these are more time-intensive to develop and require more research to understand their technical properties.[7] The SAO is developing other components of assessment within the QSAS to provide more comprehensive coverage of the standards. These include locally administered but systematically scored assessments, such as student writing portfolios and extended research projects. Given that previous research has shown that some of these components (most notably portfolios) may not be adequately objective or reliable (Gearhart and Herman, 1998; Koretz et al., 1992), the SAO is currently exploring ways in which to incorporate results from those assessments in the overall framework of the QSAS in a meaningful way.

Table 3.2 summarizes the skills assessed in the 2004 through 2007 administrations of the QCEA and the alignment process they underwent.

Item Development

For mathematics and science, no separate usability or pilot testing occurred. Instead, the field-testing of items was embedded within the operational forms of the main QCEA administration. However, for the 2004 administration, each math and science item was reviewed for cultural appropriateness by the Evaluation Institute and a team of Qatari teachers. As in the

[6] The 2006 QCEA science tests in grades 10 and 11 were expanded to include 15 standards-aligned multiple-choice items in addition to the 15 constructed-response items in the preliminary test version of the 2005 QCEA.

[7] In the spring of 2005, the SAO explored how to test oral language skills, which are part of the Arabic and English standards. The SAO decided not to include oral language skills in the QCEA because of feasibility issues and the lack of an appropriate information technology infrastructure to support computer-delivered assessments.

Table 3.1
QCEA Test Development and Alignment, 2004–2007

Development or Alignment Step	2004	2005	2006	2007
Skills assessed by QCEA	English and Arabic: reading, writing, word knowledge, grammar	English and Arabic: same as 2004 + listening	English and Arabic: same as 2005	English and Arabic: same as 2005
	Mathematics: numbers and algebra, geometry and measures, data handling	Mathematics: same as 2004 + use of manipulatives and tools + higher-order reasoning skills + piloting mental mathematics	Mathematics: same as 2005 + more data handling + piloting mental mathematics	Mathematics: same as 2006 without further piloting of mental mathematics
	Science: life science, materials, earth and space science, physical processes	Science: same as 2004 + use of manipulative and tools + higher-order reasoning skills	Science: same as 2005	Science: same as 2005
Alignment check of QCEA with standards	Postadministration check by test contractors, supervised by SAO and RAND	Preadministration check by test contractors and CfBT, supervised by SAO and RAND Alignment check by external experts	Preadministration check by test contractors, supervised by SAO	Preadministration check by test contractors, supervised by SAO
			Alignment check by external experts	

previous year, items on the 2005 QCEA for Arabic and English as a foreign language were pilot tested to ensure grade-level appropriateness and clarity and were subjected to a cultural appropriateness review by the Evaluation Institute. ETS piloted a number of constructed-response questions that required students to write sentences, short paragraphs, and long paragraphs directly onto an answer sheet alongside their answers to multiple-choice questions. ETS also piloted listening items in grades 2–12, requiring enhanced training of proctors and the development and recording of CDs. The administrative challenges of a listening component in the QCEA included finding reliable CD-playing equipment that was easy to transport yet able to produce enough volume to be audible in a classroom, as well as establishing appropriate time gaps between questions to allow students to answer. ETS also faced the challenge of finding readers to voice the listening tests for Arabic and English, given the requirement of the standards that students be exposed to a variety of accents, including those of speakers whose second language is English.

Administering the 2004 and 2005 QCEAs

Test Administration in 2004

From the beginning, efforts were made by the test companies, NORC, and the offices in the Evaluation Institute to coordinate printing and the distribution of software, data delivery files, and administration procedures. Through a series of meetings and multiparty, multisite conference calls, many accommodations were made by each party with regard to software employed,

printing schedules, administration procedures, and data delivery deadlines to ensure that the assessment data would be as secure as possible and that the processes would be as similar as possible across the four subjects to be tested.

For the 2004 QCEA, the SAO and DCMO provided practice tests to teachers and students approximately two weeks prior to the operational administration to familiarize them with the test-taking procedures and directions. In March 2004, the SEC's Communications Office launched a Web site that provided more information about the reform and answers to potential questions about the QCEA.

To ensure the security of the tests, the DCMO printed the tests on site at the Evaluation Institute. Each test company uploaded camera-ready PDF copies of each test form to a secure Web site. Once the tests were downloaded, the DCMO printing team and representatives from the test companies performed quality-assurance checks. In November 2003, the DCMO collected data on student enrollment in schools by classroom. This allowed the DCMO to assign each test to a specific student. A few extra copies of tests were printed for each classroom in case students changed classes or new students arrived in between the data collection and the test administration period. At the end of each day, any extra tests were destroyed. Hard copies of each student's tests are currently housed in a secure facility at the DCMO. The DCMO also put into place appropriate test security procedures, such as placing barcodes on each test booklet so that it could track the whereabouts of each booklet.

To maintain standardization in administrative conditions, the DCMO recruited and trained proctors on test administration procedures. The proctors were recruited to ensure a sufficient number of female proctors to serve girls' schools and male proctors to serve boys' schools. Proctors came predominantly from the expatriate community, and it was difficult to recruit male proctors.[8] To make students more at ease during the testing process, teachers were allowed to sit in the classroom during the test administration but could not assist with the administration or proctoring of the tests.

Over a four-week period in April and May 2004, proctors administered tests to all targeted schools. Secondary school students took the test during the first week of the testing period; preparatory school students, in the second week; and primary school students, in the third and fourth weeks. The mathematics and science tests included 25–30 multiple-choice items, depending on the grade level. The tests of Arabic and English as a foreign language included 16–32 multiple-choice items and one essay at selected grade levels. Testing time was 45–60 minutes for the multiple-choice component of each test. Thirty minutes were allotted for the essay portions of the tests. Students took the Arabic test on Sunday, the science test on Monday, the mathematics test on Tuesday, and the English test on Wednesday. Make-up tests were given on Thursdays.[9] If a student missed more than one test, he or she took two tests on Thursday. In this case, proctors noted on the cover sheet which make-up test the student took first. For each of the four subjects tested, 95 to 96 percent of the 88,900 students in the targeted schools sat for the test.

---

[8] Many Qataris were either employed or not looking for this kind of short-term work. Although other countries often rely on university students, retired people, or homemakers to proctor tests, these populations in Qatar did not seem interested in proctoring the exams.

[9] The school week in Qatar is Sunday through Thursday.

Test Administration in 2005

Because the 2005 administration could take advantage of the lessons learned during the 2004 administration, some changes were made to the administrative procedures. Hiring and training proctors for the 2004 QCEA proved to be expensive and logistically difficult, and there were concerns that proctors' lack of oral proficiency in English or modern standard Arabic could potentially affect students' performance in mathematics, English, or Arabic in first, second, and third grade, the questions being read aloud. For these reasons, the SAO and DCMO decided that teachers would be responsible for administering the tests in 2005. To dissuade teachers from providing answers to students during the tests, students were not proctored by their own classroom teachers but by teachers in their schools who taught a different subject. The DCMO trained teachers to be proctors in the same way it had trained external proctors the previous year.

The inclusion of the new item formats posed an additional administrative challenge because it markedly increased testing time compared to the previous year. In the 2005 QCEA, testing time in grades 1–3 ranged from 45 to 109 minutes per subject tested. Allotted testing time for grades 4–9 ranged from 120 to 160 minutes. For grades 10–12, the testing time ranged from 45 to 160 minutes, as detailed in Table 3.2.[10]

The longer testing times meant that multiple class periods were used for the administration of the tests, and this resulted in less time in which to readminister tests to students who were absent on a day of testing. Although make-ups were still available on Thursdays during

Table 3.2
2005 QCEA Testing Times, by Subject and Grade

	Subject (minutes)			
Grade	Arabic	English	Mathematics	Science
1	89	93	45	Not tested
2	100	109	45	Not tested
3	100	108	45	Not tested
4	127	121	160	120
5	127	133	160	120
6	135	145	160	120
7	135	145	160	120
8	135	140	160	120
9	135	149	160	120
10	160	149	80	45
11	160	149	80	45
12	160	149	80	45

[10] Mathematics and science tests in grades 10–12 were shorter than those in grades 4–9 because they were only preliminary versions of more fully aligned tests. They contained 15 constructed-response items intended to measure multiple standards.

the test window, it was decided that make-up tests would not be administered in listening because of the complications of administering different listening tests in separate test rooms in one day. In addition, because of the increased length of the tests, students made up only one test, even if they missed more than one.

The decision to hold make-up tests for the QCEA in 2004 and 2005 and the procedures that the make-up testing would follow were a compromise among a number of factors. The foremost concern for the SAO was capturing a large portion of the targeted student population to make valid inferences from the assessments. Of primary concern for the DCMO was the feasibility of scheduling make-up tests. Added to the possibly divergent priorities of each office was the fact that no information existed prior to the 2004 administration on absentee rates or on how absent students differed from the broader testing population. It was therefore difficult for either office to speculate on the effects that a make-up test would have on the test validity or on how administratively difficult employing a make-up test policy would be. Although the Evaluation Institute decided not to hold make-up tests for the 2006 QCEA, it is vital for the success of the testing operation that the Evaluation Institute understand how best to capture as many students as possible without overburdening the system or compromising standardized conditions, while emphasizing to the public the importance of the tests.

CHAPTER FOUR

Scoring the QCEA and Reporting Results

A favorable public perception of the assessment operation is central to the ENE reform's success. Without public support and approval, parents, teachers, students, and school administrators will not trust results and may even lose faith in the reform effort itself. With that in mind, the SAO and RAND teams worked diligently with the test developers to craft reporting mechanisms that would be meaningful and understandable to the public at large. This chapter documents the scoring and reporting of results of the 2004 and 2005 QCEAs. We also compare results from the 2005 and 2006 QCEAs.

Scoring the Tests and Reporting the Results from the 2004 QCEA

From June through August 2004, NORC and the DCMO electronically recorded multiple-choice responses at an on-site computer facility, and electronic records of the students' responses were forwarded to the testing contractors for scoring. It took longer than anticipated to produce these data, primarily because of NORC and the DCMO's inexperience in administering and compiling the results of a testing operation, particularly one of this magnitude. This inexperience unfortunately resulted in a number of anomalies in data handling.[1] Multiple-choice scores from the test contractors were provided to the SAO as percent correct in November 2004, with scale scores provided shortly thereafter. The Arabic and English essay questions were scored by 217 Arabic and 57 English teachers in Qatar who had been trained by ETS. The essays were rated on a seven-point rubric scale (0–6) and reported separately from the multiple-choice results.

The Evaluation Institute carefully considered how best to report results to the public. One challenge was to provide accurate information to the public that would not invite inappropriate or inaccurate comparisons between the 2004 and future QCEA administrations. After numerous meetings in the fall of 2004 among the SAO and RAND teams and the test companies, a communication plan was developed that would present general, skill-level (e.g., reading comprehension, algebraic computation) results by grade for students from different subgroup populations. Those subgroup populations would be Qatari and non-Qatari, boys and girls, and Ministry and private Arabic schools.[2]

[1] For example, some students had been assigned to one classroom but had taken their tests in another classroom and, therefore, had been treated as "missing" in the database, QNEDS. Because efforts to recapture these students' scores were time-consuming and labor-intensive, the data were not fully available to ETS and CTB until September 2004. Up to that point, the test developers had been working with data sets that were 98-percent complete.

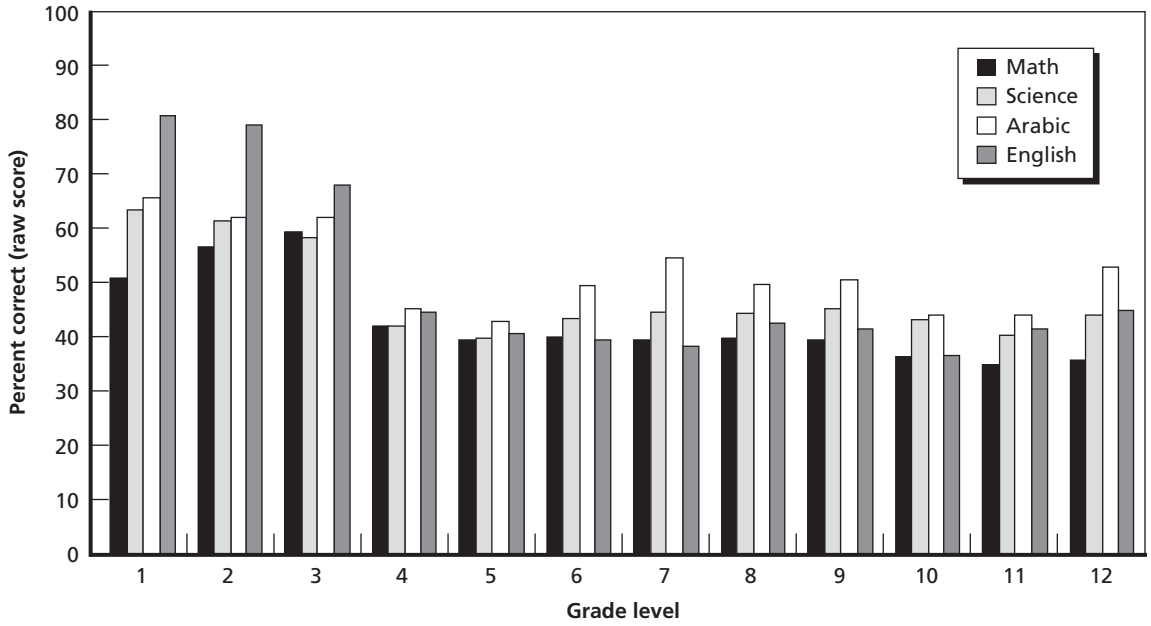
[2] Independent schools had not yet opened when the 2004 QCEA was administered.

The director of the Evaluation Institute presented the skill-level results alongside illustrative example items during a March 15, 2005, public discussion meeting, or *hiwar*.[3] In all subjects and in most grades, girls outperformed boys. In all four subjects and in all grades, non-Qatari students outperformed Qatari students. The comparison between Ministry schools and private Arabic schools showed mixed results for each subject and grade.[4]

Figure 4.1 displays the 2004 QCEA results for students in Qatar's Ministry of Education schools and private Arabic schools by grade and subject in terms of the percent of correct test items on the multiple-choice component of the assessment.

On average, students in grades 4–12 answered around 40 percent of the test questions correctly. Specifically, students in 11th and 12th grades correctly answered, on average, only 35 percent of the mathematics questions and around 40 percent of the science and English questions.[5] These data indicate that students may not have been graduating secondary school with a knowledge base consistent with international standards, as defined by the test contractors (described in Chapter Two). On the other hand, the proportion of correct responses was much higher in grades 1–3. This may have been because the test administration procedures

Figure 4.1
Percent Correct, QCEA Multiple-Choice Questions, 2004



SOURCE: Data from Qatar Supreme Education Council.

RAND *TR620-4.1*

[3] *Hiwar* is Arabic for *discussion*. On March 15, 2005, the Communications Office of the SEC organized several panel discussions at which expert speakers talked about Qatar's education reform. One session was devoted to an update on the results of the assessments and surveys administered in 2004. The presentation is not publicly accessible.

[4] Scale scores and percent correct for each subject and each student are available in the database system, QNEDS, and can be accessed with permission from the assistant director of the Evaluation Institute. Item response scores were not available as of this writing.

[5] These results should be interpreted with caution. The statistics are raw, unequated student scores, and therefore no between-grade or cross-subject comparisons should be made.

in grades 1–3 differed from those in grades 4–12: In the early grades, instructions and test items were read aloud to students. These results supported general perceptions of student performance found in previous studies (Brewer et al., 2007).

Scoring the Tests and Reporting the Results from the 2005 QCEA

Similar to the 2004 QCEA, the multiple-choice questions on the 2005 QCEA were scored by computer, and the Arabic and English essay questions were scored on a seven-point rubric scale by Arabic and English teachers in Qatar who had been trained by ETS. The constructed-response questions in all four subjects were also scored by carefully selected teachers and other suitable personnel who were trained by the respective testing company. Scoring occurred in Qatar over a six- to 12-week period directly after the administration.

In February 2006, the Evaluation Institute communicated national-level results of the 2005 QCEA to the public and described how well students were performing relative to Qatar's new curriculum standards by grade, school type (Ministry of Education, Independent, or private Arabic), and gender. The SAO delineated three main performance levels to describe QCEA results:

1. *Meets standards:* The student has completely fulfilled expectations for the acquisition of knowledge and skills required at his or her grade level by the Qatar curriculum standards.
2. *Approaches standards:* The student has fulfilled the minimum expectations.
3. *Below standards:* The student has failed to fulfill the minimum expectations. This performance level is divided into three sublevels (3, 2, and 1). Because the below-standards level was by far the largest category, the SAO thought that it was important to subdivide the performance level further to enable it to report the progress students were making within the larger category. The sublevels were named according to the degree of effort necessary to move to the next category:
   - Level 3: The student may reach the approaches-standards level with some additional effort.
   - Level 2: The student may reach the approaches-standards level with considerable additional effort.
   - Level 1: The student may reach the approaches-standards level with extensive additional effort.

Table 4.1 uses fourth-grade mathematics as an example of the types of expectations for student performance at each level.

As explained in Chapter Three, an external audit of the alignment between the 2005 QCEA and the Qatar curriculum standards found that each item on the 2005 QCEA matched a standard but that the test as a whole did not provide adequate coverage of the breadth and depth of the skills and content dictated by the standards. The 2005 QCEA was therefore only "partially" aligned to the standards.[6] Although partial alignment was less than ideal, the SAO determined that the public and political demand for performance levels necessitated

[6] This was to be expected, given that the standards were not finalized until after the 2005 QCEA was developed.

Table 4.1
Student Performance-Level Expectations, Grade 4 Mathematics

Performance Level	Proficiency
Meets standards	Able to apply knowledge of decimals, fractions, and factors Able to explain basic geometry relationships (e.g., parallel/perpendicular, rectangle/square, area/perimeter) and use these concepts to show and explain the work involved in solving problems Able to read and interpret graphs and understand the relationship between graphs representing the same data Given a rule for a pattern, able to apply that rule to solve problems Able to represent a mathematical situation with a simple algebraic equation and apply that equation to solve a word problem
Approaches standards	Able to understand basic geometry definitions (e.g., right angle, parallel, perpendicular, line of symmetry) Able to multiply whole numbers and multiply/divide by 10, 100, and 1,000 Able to round small numbers and estimate Able to solve simple one-step word problems Able to represent a simple situation using fractions and identify equivalent fractions
Below standards	Has not acquired the basic knowledge and skills required for this grade level as measured by the Qatar curriculum standards Can read and write whole numbers, round simple decimals, and identify angles and lines

their development. Performance categories provide concrete descriptions of the types of skills and knowledge expected of students falling into a particular category, which can help education policymakers, school administrators, teachers, parents, and students interpret the scores in a meaningful way and understand how well students in Qatar are performing vis-à-vis a set of standards.

To determine the performance categories, a panel of content experts in each subject is typically convened. After discussing and reaching consensus about the skills and knowledge that are representative of each category, the panel members make independent judgments regarding the placement of cutoff points, or cut scores (the bookmark method), or they rate the probability that different groups could solve different items correctly (the modified Angoff method).[7] Together, both methods enable the establishment of different performance categories that describe what students in each category know and are able to do. The SAO convened local educators to define the performance levels. Experts from CTB and ETS led the workshops and assisted with the psychometric development of cut scores. The Evaluation Institute made the final decision regarding the placement of cut scores. In April 2006, the Evaluation Institute's SEO disseminated school report cards noting the 2005 QCEA scale scores at the grade level for each subject by school alongside other school-level indicators taken from the surveys administered in 2005.[8]

Scoring the Tests and Reporting the Results from the 2006 QCEA

The Evaluation Institute reported results for the 2004 QCEA at the national level and for the 2005 QCEA at the national and school levels. For the 2006 QCEA, the Evaluation Institute

[7] Cut scores for the Arabic and English performance levels were determined with the modified Angoff method. Cut scores for the mathematics and science performance levels were determined with the bookmark method.

[8] School report cards can be found in Arabic on the SEC's Web site (see Qatar Supreme Education Council, 2005).

reported results at the national, school, classroom, and individual student levels. In March 2007, the SAO developed reports for students and their parents with performance level, scale scores and information on how the students could improve their test results in all four subject areas. The reports included information about the QCEA and how to interpret the test results. At that time, the SAO also disseminated classroom reports intended to provide the teachers with an impression of the proficiency of students under the previous year's teachers. This information also gives the teacher a better understanding of different proficiency levels in relation to the curriculum standards and the need for remediation. Apart from providing feedback about where students stand vis-à-vis the standards, this information enables teachers in Independent schools to better focus instruction for groups of students with different levels of proficiency.

Table 4.2 summarizes the reporting mechanisms from 2004 to 2007.

Comparing 2005 and 2006 QCEA Results by School Type

It has been only a few years since the inception of the reform, and it is therefore difficult to ascertain the extent to which the reform is meeting its goals or fostering other changes in the education system. Table 4.3 shows the percentage of students in each school (Independent, Ministry of Education, and private Arabic) who are meeting the standards or approaching the standards in each subject area in grades 4, 8, and 11. (Results for all grades are available in Appendix C.) In each subject and in all grades assessed, Independent school students performed better overall than did students from Ministry or private Arabic schools, particularly in English language. However, only a small percentage of students from any school are meeting the standards. In fact, no students from any of the schools are meeting the standards in science, and very few are meeting them in mathematics. This is true even among Independent school students. This result is, however, not too surprising given the challenging nature of the curriculum standards, the time needed to develop curricula to teach the standards, and the fact

Table 4.2
QCEA Proficiency Levels and Reporting of Results, 2004–2007

Element	2004	2005	2006	2007
Standard setting in QCEA (proficiency levels)	None	Arabic and English: modified Angoff method	Arabic and English: used cut scores established in 2005	Arabic and English: same as 2006
		Mathematics and science: bookmark method	Mathematics: used cut scores established in 2005	Mathematics: same as 2006
			Science: used cut scores established in 2005 for grades 4–9; resetting of performance levels for grades 10 and 11 due to changes in test blueprint	Science: same as 2006 (no resetting)
Level at which QCEA results were reported	National	National School	National School Classroom Individual	Same as 2006

Table 4.3
Performance-Level Results of 2005 and 2006 QCEAs, by Subject and School Type, Grades 4, 8, and 11 (percentage distribution)

Subject and School Type	Grade 4				Grade 8				Grade 11			
	Meets Standards		Approaches Standards		Meets Standards		Approaches Standards		Meets Standards		Approaches Standards	
	2005	2006	2005	2006	2005	2006	2005	2006	2005	2006	2005	2006
Arabic												
Independent	6	3	40	31	12	4	34	30	19	9	49	35
Ministry of Education	3	2	19	21	4	3	20	20	4	4	22	24
Private Arabic	3	1	22	21	5	3	22	18	3	3	29	18
English												
Independent	10	7	32	21	11	5	27	14	23	13	39	24
Ministry of Education	0	0	3	3	1	1	6	5	2	1	7	8
Private Arabic	2	0	9	4	2	2	12	8	6	1	10	9
Mathematics												
Independent	0	0	33	41	0	1	51	49	0	0	69	48
Ministry of Education	0	0	17	17	0	0	17	15	0	0	23	25
Private Arabic	0	0	19	16	0	0	19	16	0	0	29	28
Science												
Independent	0	0	41	16	1	0	49	35	0	0	48	43
Ministry of Education	0	0	17	14	0	0	15	14	0	0	25	22
Private Arabic	0	0	16	13	0	0	16	11	0	0	28	25

that students have only recently been exposed to those standards. Furthermore, across the subjects, students' performance on the QCEA, in general, declined between 2005 and 2006.

Table 4.3 does not disaggregate the data by an Independent school's year of opening, student gender, or student nationality. QCEA results in 2005 and 2006 reveal that girls tend to outperform boys and that non-Qataris tend to outperform Qataris at most grade levels in each subject.[9]

The remainder of this chapter summarizes the results by subject area.

Arabic and English

On the English and Arabic assessments in 2005 and 2006, a greater percentage of Independent school students were meeting or approaching the requirements set by the curriculum standards than were students in either Ministry of Education or private Arabic schools. The difference among schools is most striking in 2005, the end of the academic year in which the first cohort of Independent schools opened. Six percent and 10 percent of fourth-grade students in Independent schools met the Arabic and English standards in 2005, respectively. Whereas only 3 percent of students in Ministry and private Arabic schools met the Arabic standards, no students in Ministry schools met the English standards, and only 2 percent of students in private Arabic schools met the English standards.

The differences are more apparent in the higher grades. Twelve percent and 19 percent of Independent school students in grades 8 and 11, respectively, met the Arabic standards. Only 4 percent of students in Ministry schools in grade 8 or 11 met the Arabic standards. Five percent and 3 percent of students in private Arabic schools met the Arabic standards in grades 8 and 11, respectively. In English, 1 percent of eighth-grade students in Ministry schools met the standards, and 2 percent in 11th grade did so.

Although the Independent school students still out performed their Ministry of Education and private Arabic school counterparts in 2006, the numbers of students who were meeting or approaching the standards in Independent schools dropped. This is particularly notable in grade 11. In 2005, 19 percent of 11th-grade students in Independent schools met the standards in Arabic, whereas only 9 percent of Independent schools students in 11th grade met the standards in 2006. In 2006, the total percentage of 11th-grade students approaching or meeting the standards dropped, with only 5 percent meeting the standards. A similar pattern emerges for the English tests. In 2005, 23 percent of 11th graders in Independent schools met the standards. In 2006, that number dropped to 13 percent.

Mathematics and Science

Except for very few exceptions, students in Qatar are far from meeting the skill levels needed to reach the meets-standards performance category in mathematics and science. In 2005, a larger percentage of students in Independent schools were approaching the standards in mathematics and science relative to students in Ministry of Education or private Arabic schools. The same was still generally true in 2006, but the difference between the school types had shrunk as more schools were converted from Ministry to Independent schools.

In 2005, only a handful of Independent school students in ninth and 10th grades met the mathematics standards (see Appendix C); the same was true among eighth graders for the

[9] QCEA results by gender, subject, and school type for selected grades for 2004 (in Arabic), 2005, 2006, 2007, and 2008 can be found on the SEC's Web site (see Qatar Supreme Education Council, undated[b]).

science standards. About 1 percent of Independent school 11th graders met the standards in 2006. However, if we compare the general trends of students approaching the standards, it is clear that, as in the Arabic and English results, Independent school students perform better than students in Ministry or private Arabic schools but that the relative difference was less extreme in 2006.

CHAPTER FIVE

Lessons Learned and Future Directions

As we reflect on the early years of test planning, development, administration, and reporting for the QSAS and the QCEA, a number of important lessons emerge. These lessons can provide helpful guidance for education policymakers around the world who are looking to implement standards-based assessments in their education systems. This chapter concludes with recommendations for the Evaluation Institute as it continues to implement the SAO and RAND teams' vision for the QSAS.

Lessons Learned from Developing and Implementing the QSAS and QCEA

Separation of Standards Development and Assessment Development Hampered Communication Around Alignment

As shown in Figure 1.1 in Chapter One, the initial design of the Education and Evaluation Institutes placed responsibility for developing the standards with one entity, the CSO in the Education Institute, and responsibility for developing the assessments with another entity, the SAO in the Evaluation Institute. In theory, these two offices were to work together to develop a testing system that was fully aligned with Qatar's national curriculum standards. However, in practice, because there was not an explicit structural connection and, therefore, no formal feedback mechanism or liaison between the two offices, the two did not always work in tandem. Adding to this problem was that the CSO did not have a permanent office director until June 2004 to establish informal or formal connections with the SAO, nor did it have curriculum specialists in all four curriculum areas in the early years of test development. These staffing problems further curtailed consistent or coherent communication between the offices. Instead, the SAO had to rely on the standards-development contractor, CfBT, to provide feedback on the links between the test items and the standards. In the past year, the assistant director of the Evaluation Institute has taken steps to develop a joint commission to formalize the relationship between the two offices but has not been able to make much progress to date.

For education policymakers considering implementing a standards-based assessment system, we recommend that formal linkages be built between standards-development and assessment-development authorities prior to implementation. This could be a permanent staff member with explicit duties to liaison between the two offices, or the curriculum staff and assessment-development staff could be housed within the same office.

The Timeline for Developing a Fully Aligned Standards-Based Assessment System Was Too Short

Education policymakers in Qatar felt the need to jump-start the reform with a comprehensive standards-based assessment. The first wave of Independent schools opened their doors in September 2004. The expectation was to have a standards-based assessment system in place by the end of the 2004–2005 academic year so that students' learning vis-à-vis the newly developed curriculum standards could be appropriately measured. This timeline gave the SAO, the RAND team, and the test developers approximately 18 months from the time ETS and CTB were selected in May 2003 to January 2005, when final proofs of the exams had to be ready for printing. There were a number of challenges in meeting this deadline: By 2005, the QSAS's goals, purposes, uses, and design features were laid out, but the SAO and RAND teams were unable to finalize a detailed blueprint or implement the system's features by this date. There were three reasons for this delay. First, as described in Chapter Three, the SAO and RAND teams decided to focus efforts on developing the core component of the QSAS—the QCEA—which was to be the largest and most comprehensive component of the system. Other summative assessments designed as components of the QSAS were to measure skills and knowledge not adequately assessed by the QCEA. Given the size and scope of the QCEA and the fact that the SAO would need to coordinate the delivery of any other assessments to Independent schools, which had not yet opened, it made sense to start with the QCEA and then, over time, develop the other components. Second, in 2003 and 2004, the SAO had only three staff members, which limited the office's capacity to focus on the implementation of the QCEA alongside the implementation of other components of the QSAS. Third, the SAO and RAND teams and the test developers worked with draft curriculum standards until they were finalized in 2005; final decisions on the QSAS design could not occur until the standards were completed.

To meet the leadership's timetable for K–12 education reform, the SAO started the process of aligning the QCEA with the curriculum standards in December 2003. The final versions of the standards were available for mathematics, science, and English in January 2005. As explained in Chapter Three, the timeline was particularly challenging for the development of the Arabic tests because the Arabic standards went through numerous iterations and the final version was not delivered to the CSO until May 2005. This issue also affected the development of the 2005 QCEA: Test development and necessary pilot and usability studies occurred in a less-than-ideal timeline. For example, the test items were piloted concurrently with the main administration. Because of this, more questions were given to students than otherwise would have been, in case questions did not "work" as anticipated.

It was particularly challenging for the SAO to develop ideas for an assessment system when responsibilities for implementing the different components of that system were not yet clearly determined. For example, questions remained about whether the Education Institute or the Evaluation Institute would develop and administer formative assessments that were intended to be part of the QSAS. Furthermore, given the CSO's staffing challenges, discussed earlier, and the fluctuation of key decisionmakers within the Education Institute, it was difficult to map out a plan for an assessment system for Qatar early on because key players in the Education Institute were not yet hired or had changed rapidly.

For education policymakers considering implementing a standards-based assessment system, we recommend at least three years to develop a comprehensive, fully aligned standards-based assessment system, as suggested by experts (Commission on Instructionally Supportive Assessment, 2001; Pellegrino, Chudowsky, and Glaser, 2001). We suggest even more time

if performance-based assessments are to be applied. Three years would allow adequate time for test developers to build test blueprints, create items based on the standards, pilot items, field the tests, and make adjustments to the assessments if and when standards are modified. For education systems that may encounter similar staff challenges and the possibility of rapid policy shifts—as experienced in Qatar—we recommend five years.

Logistic and Administrative Constraints Often Took Precedence Over Substantive Needs of the QCEA Testing Operation

The Evaluation Institute was newly established and members of the DCMO had no experience with delivering, coding, or managing a testing operation of the size and scope of the QCEA. Furthermore, school administrators, teachers, parents, students, and even proctors had no experience with the demands of a testing operation of this scale. This inexperience led to a number of decisions that gave priority to logistical issues over substantive ones. Important decisions, such as testing time, delivery date of the final test booklets to the printers, administering make-up tests, and who would proctor the QCEA, were predicated on logistical concerns rather the substantive concerns of research staff on the SAO and RAND teams. For example, as noted in Chapter Three, pilot tests in 2003 were limited to 45 minutes (the length of one class period) to avoid inconveniencing students or teachers. Although not wanting to overburden students or teachers should certainly be a high priority, the fact that the tests had to be shortened to fit into the allotted time required that more classrooms be tested than if the test length were longer. An unforeseen consequence of limiting test time was that, by the time proctors finished instructing the students on how to take the mathematics tests, only limited class time remained for actually taking the test; this resulted in many nonresponses, thus limiting the ability of CTB to analyze the mathematics items' usability. The test time for the 2004 QCEA was subsequently lengthened based on the problems encountered in the 2003 usability test.

For education policymakers considering implementing a standards-based assessment system, we recommend that prior to the administration of a test, the entities in charge of developing and administering the tests agree on administration processes and procedures that strike a balance between limiting student burden or fatigue and ensuring that appropriate analyses can be made from the test results.

Many Policies About Testing Did Not Consider Existing Research or Analysis

Given how rapidly the reform's implementation occurred, many policy decisions had to be made early on without giving necessary consideration to research findings or without appropriate guidance to schools. One example is the decision in 2005 to have Independent schools move toward teaching mathematics and science in English prior to developing a coherent language-of-education policy that would have provided the nation with specific goals vis-à-vis English and Arabic language learning.

Another example is that the 2006 QCEA was offered in both Arabic and English to accommodate growing concerns about perceived disadvantages among students who were instructed in English but tested in Arabic. As mentioned earlier, the RAND team designed, and the SAO executed, an experimental study during the 2005 QCEA for fifth graders who were taught mathematics and science in English. Although the SAO did not find much evidence to support the hypothesis that these students would be at a disadvantage if they took the QCEA in Arabic, a policy decision was made to offer the 2006 QCEA in both Arabic

and English, depending on a school's choice. While having two forms would allow the SAO to further study the effects of taking tests in one's language of instruction rather than in one's mother tongue, the point remains that this decision was made before the results from the 2005 administration had been fully analyzed. A third example is the decision in November 2006 to limit testing to only those students enrolled in the Independent schools. As noted, the decision had a number of potential drawbacks, which may not have been fully considered.

For education policymakers considering implementing a standards-based assessment system and for education policymakers in Qatar, we recommend that research findings and empirical evidence be considered in decisionmaking. If the Evaluation Institute, the Education Institute, and the governing body of the SEC are to make informed policy decisions about the assessments and student achievement, they must base those decisions on empirical evidence, lest innuendo or unfounded perceptions sway education policy in the nation.

There Was Insufficient Communication About the Purposes and Uses of Testing

Many questions arose from the public over the years about the purpose of the QCEA and its implications for students in Qatar's schools. Close to the 2004 and 2005 testing operations, the Evaluation Institute was flooded with phone calls from parents, teachers, and administrators. Furthermore, editorials in local newspapers called into question the need for a testing operation of such length. Although the SEC has a Web site with a list of frequently asked questions, and the March 2005 and 2006 *hiwar* forums presented national-level assessment results, there was no concerted effort to communicate directly to stakeholders who may not have had computers, Internet connections, or the time to attend a forum about the QSAS in the middle of the day. The phone calls and editorials in the newspapers suggested that the public perceived the Evaluation Institute's lack of communication with the public as a lack of direction or a lack of purpose of the QCEA. This perception, we would contend, may have lowered the motivation for students to do well on the QCEA and for teachers to support the QCEA testing.

For education policymakers considering implementing a standards-based assessment system, this problem could be addressed early on through direct outreach to a number of constituents. Sessions for parents and other community stakeholders might be scheduled over weekends or in evenings during the week, when working adults can attend meetings. (For Qatar, evening meetings would be the most appropriate.) Meetings with other education stakeholders (for Qatar, these include Independent school operators, teachers, and Ministry of Education personnel) might occur on a continuous basis throughout the early years of testing. Although the SAO met with Independent school operators to explain the purposes of the testing system, no systematic outreach effort was extended to other Independent school administrators or to educators in the Ministry of Education.

Furthermore, public acceptance of the assessment system could have been enhanced by improving the transparency of the testing operation. In other testing operations, this problem could be addressed early on by providing individual-level achievement data from the first year of testing. (For the QCEA, individual-level data were available only after the third year of testing.)

Challenges That the Evaluation Institute Should Address

The QSAS has made significant progress toward the goals of the initial design of using multiple methods to assess student progress relative to the standards, yet additional areas still need to be addressed. As the QSAS continues to evolve, the Evaluation Institute and the SEC will need to attend carefully to some future challenges that will bear on the alignment of the QSAS to the standards and the future of standards-based education reform in Qatar. We discuss each in turn.

Assess Content from the Advanced Standards

The standards for secondary school students are divided into foundation and advanced levels, and the QCEA now tests foundation standards only. As students become increasingly prepared for, and enrolled in, the advanced track, the QSAS will need to include content that is fully aligned to the advanced standards in order to gauge the performance of some of Qatar's most promising students. Given the low percentage of students who currently meet the standards on the QCEA, this issue may not present itself for a number of years. However, in 2007, the Education Institute established a policy that all students in Independent schools must take at least two courses that follow the advanced standards. Because of this new policy, the Evaluation Institute will face this issue sooner than expected and should start developing questions now that are aligned with the advanced standards.

Provide Accommodations or Alternative Assessments for Students with Disabilities

Students with disabilities are currently excluded from the QCEA for a variety of reasons, including the lack of data about the percentage of Qatari students who need accommodations and the types of accommodations that should be provided. Further complicating the issue of inclusion is whether the skills and concepts specified in the standards are appropriate, especially for students with significant disabilities (Roach, Elliott, and Webb, 2003). However, it is important to test these students to the extent possible because the overall effectiveness of the reform cannot be measured without considering the performance of all students. Future policy debate should focus on ensuring that students with disabilities have opportunities and instruction that allow them to make progress toward the standards, where deemed appropriate. To incorporate these students into the QSAS, the Evaluation Institute will need to develop accommodations for disabilities on the tests.

Use More Advanced Technologies

Another promising avenue that may allow for more comprehensive coverage of the standards is the use of computers to administer the tests, as discussed earlier. With tests delivered through computers, graphics, sound, running video, and text can be combined to present items that require examinees to demonstrate a variety of skills that cannot be easily measured with paper-and-pencil tests. Some of these tasks include performance-based items that measure students' ability to carry out research with auxiliary reference materials and tasks that assess listening proficiency and oral communication. The Evaluation Institute should explore the use of more advanced technologies to augment the QCEA, as well as whether computer-based delivery of assessments is feasible given the country's information technology infrastructure.

Communicate with the Public

The Evaluation Institute has already taken some steps to communicate with the public in yearly open forums and through the school report cards, first published in April 2006 using 2005 QCEA data, which contain a variety of other information about schools, including staff, facilities, and programs, that will inform parents' decisions about where to send their children. In addition to the school-level results available on the school report cards, the SAO disseminated results from the 2006 QCEA at the national, classroom, and individual levels in March 2007. To enhance the utility of the QSAS to inform policy, the Evaluation Institute should consider a variety of strategies to widely disseminate the test results. It could hold press conferences and other public events to address questions about the QSAS or publish research- or practitioner-oriented reports.

Conduct Validity Studies

Of vital importance is the need for validity studies of the QCEA—continuous studies that test the soundness of interpretations of the test scores for a particular use. According to leading experts in test validity (see Cronbach, 1988; Messick, 1989; Shepard, 1993; Kane, 1992, 2006; and American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999), validating a test is an ongoing process. Any number of sources of evidence can be incorporated to develop a coherent and comprehensive argument to support the QCEA's uses.

Finalize Policy Decisions in Designing Future QSAS Administrations

There are a number of design features that need to be implemented and policy decisions that need to be finalized as the QSAS moves forward. One important feature that needs clarity is how best to implement formative assessments in the classroom, to assist teachers in knowing how well their students are learning the standards. Another issue is the development of creative assessments that test scientific inquiry or hands-on processes. As the technological infrastructure in Qatar improves, offering computer-based assessments may ease the application process of these types of assessments. These decisions need to be finalized and communicated with the public to ensure the success of the assessment system.

Concluding Thoughts

The development of the QSAS was an extensive undertaking that is still ongoing. First fielded in April and May 2004, the QCEA represented the first time that students in all grades in government-sponsored schools in Qatar were tested in a systematic, standardized way. The questions were presented in a multiple-choice format with essays for students in upper grades. Although the QCEA in 2004 was not aligned to the curriculum standards, it nonetheless provided preliminary information about what students know and can do. Assessing students' skills in English, Arabic, mathematics, and science provides one set of indicators to evaluate the extent to which students have the skills they need to succeed in further education and be productive members of the Qatari society and economy. Around 88,000 students in Ministry and private Arabic schools participated in the 2004 QCEA, or approximately 95 percent of the targeted student population.

The second QCEA, administered in April and May 2005, was designed to align with the curriculum standards and included a variety of test formats, including multiple-choice, constructed-response, essays, and performance-based items. One of the cornerstones of standards-based reform is aligning the tests to the standards. It is important to recognize that alignment requires careful attention at various stages of the standards- and assessment-development processes. Standards are not a static set of competency statements: They reflect the goals of the education system and society at large at the time they are written. As societal trends change, so will the expectations of knowledge and skills required of students. It is imperative that the curriculum standards and assessments reflect these changes over time. Standards provide a framework in which to develop assessments, which, in turn, provide information about the attainment of the standards. Information about how well students master the content standards can prompt instructional or curricular improvements. In this way, alignment plays an important role in improving teaching and learning relative to the standards (Long and Benson, 1998).

Although the current QSAS has implemented many of the initial design features, the assessment system is not yet complete. Comprehensive coverage of the curriculum standards will not likely be possible without the expansion and diversification of testing alternatives beyond the QCEA. Currently, the SAO is exploring the feasibility of locally administered measures, such as portfolios. Additionally, the SAO is evaluating the capacity to deliver the QCEA through the computer, which would allow for shorter testing time and innovative item formats that are not supported by conventional paper-and-pencil measures. While these features cannot be immediately implemented, they represent promising ways of improving the QSAS.

The ENE reform initiative substantially changed the testing landscape in Qatar. Prior to the reform, there was a dearth of systematic and objective information on students' achievement and skills. This changed with the advent of the QCEA, which represents Qatar's first standardized testing system for students in government-sponsored schools. This was no easy feat, and it was managed by a small group of Evaluation Institute staff. With its experience in testing, evaluation, and information technology, this group forged relationships with contractors across the Middle East, Europe, and the United States to successfully develop, administer, warehouse, and report results from the new standards-based assessment. It was the first time an operation of this scope and depth has occurred in the region.

APPENDIX A

Assessment Elements Considered for the QSAS

Table A.1 lists the variety of components that the SAO and RAND teams considered for the QSAS. Many of the components are still under discussion within the SAO.

Table A.1
Components Considered for the QSAS

Component	Item Type	Benefits	Drawbacks
Conventional paper-and-pencil tests	Multiple-choice	Well-established, relatively cost-effective methods for development Strong psychometric properties Not as disruptive to instruction as computer-adaptive tests because all students can take the test simultaneously	Do not estimate student proficiency as well as CAT or level tests, especially at the lower and upper ends of the proficiency scale Require more testing time than do level or computer-adaptive tests to obtain same level of precision Because items are not tailored to student proficiency, the tests can be "demotivating," as items may be much too difficult or too easy for students' proficiency level
Paper-and-pencil level tests	Multiple-choice	Same advantages as conventional paper-and-pencil tests Allow for greater accuracy in estimates of student proficiency than with traditional paper-and-pencil tests Require shorter testing time than do conventional paper-and-pencil tests because students are not being tested on items that are too difficult or too easy	Although these tests provide better estimates of student proficiency than conventional paper-and-pencil tests, they are not as accurate as CAT Shortened test length makes alignment check with standards difficult Lower face validity due to perceived lack of fairness to students Lower-performing students exposed to only less demanding questions and not to more challenging questions may not understand full range of expectations in the course
CAT	Multiple-choice (though may also include open-ended items that do not require human scoring)	Allow for greater accuracy in estimates of student proficiency in shorter periods of testing time than conventional paper-and-pencil tests because it eliminates items that are too difficult or too easy for an examinee Provide immediate feedback of scores to students and teachers Enhance test security relative to conventional paper-and-pencil tests because questions are drawn from an item bank consisting of thousands of possible items, thereby decreasing the risk of students being exposed to the items prior to testing Decrease "teaching to the test" because teachers are less likely to be able to coach individual students on the specific content they will encounter Broaden the range of skills and knowledge that can be measured while still offering efficient scoring (e.g., fill in short answers, move items around the screen, create a food chain)	More costly than paper-and-pencil tests Do not allow for complex open-ended types of questions, which can limit the test's validity Logistically more difficult to implement than paper-and-pencil tests because it is unlikely that all students can be tested simultaneously Most vendors have little or no experience administering computerized tests on a large-scale basis Although they may improve test security in some ways, they create other security concerns (e.g., remote access to item pools) Same concerns as for level tests, plus development can be time-consuming

Table A.1—Continued

Component	Item Type	Benefits	Drawbacks
Performance assessments	Math: open-ended problem-solving items Writing (in English and Arabic): essays in different genres (e.g., narrative, persuasive, informative) Science: hands-on or open-ended	Open-ended tests can assess skills that are not amenable to measurement via multiple-choice tests; for example, an essay test is arguably a better measure of students' writing proficiency than a multiple-choice test, which is more likely to assess a different aspect of writing—namely, editing skills More similar to the kinds of teaching that the standards aim to promote More likely than multiple-choice tests to create incentives for teachers to include writing and problem solving in their instruction	More costly than multiple-choice tests to develop and score Tend to have poorer psychometric properties per unit of testing time than do multiple-choice tests Scoring mechanism (human raters) introduces an additional source of bias and variability in the scoring process
Portfolios	Open-ended	Can positively affect instructional practices (e.g., teachers more likely to include problem solving in mathematics instruction) Less intrusive than externally mandated tests because the classroom work samples are an inherent part of the curricula	Poor psychometric properties, particularly low objectivity and low reliability Extremely resource-intensive (in terms of both time and money) to develop and score Scoring mechanism (human raters) introduces an additional source of bias and variability to the scoring process

APPENDIX B

Steps to Align Assessments with Curriculum Standards

This appendix provides a detailed overview of the QCEA alignment process, starting with a theoretical discussion of the concept of alignment.

The Concept of Alignment

Alignment refers to the extent to which standards, curriculum materials, assessments, and other elements of the education system work together to guide instruction and student learning (Webb, 1997). Implicit in the push for alignment is the assumption that a coherent system will provide teachers and students with clear signals about what is important to teach and learn (Smith and O'Day, 1990). Especially with a standards-based reform system, establishing a high degree of alignment between test content and standards is a key piece of validity evidence supporting the use of the test scores to inform decisions about whether students have mastered the skills and content specified in the standards (Lane, 1999).

Currently, there is no consensus about best practices in terms of how to examine alignment between test content and standards; criteria about what constitutes "sufficient" alignment vary from study to study, as do the methodologies used to evaluate alignment. In a review of the literature, La Marca, Redfield, and Winter (2000) identified two overarching dimensions found in most alignment studies. The first dimension, referred to as *content match*, pertains to the degree to which the test content is congruent with the content in the standards. This dimension addresses the extent to which the test assesses specific objectives or indicators and the extent to which its coverage matches the emphasis as designated in the standards (La Marca, 2001). The second dimension, referred to as *depth match*, pertains to the match between the cognitive complexity of the test items and the cognitive complexity prescribed by the standards.

Achieving alignment is a multistage process that requires extensive reviews of both the standards and the test items (Webb, 1997, 1999). Typically, a panel of content experts is convened to systematically review the standards for their amenability to measurement. Many standards can be worthwhile (e.g., "students develop an appreciation for mathematics") but written in a manner that is too general or diffuse to be measured (Ananda, 2003). Once the standards have been reviewed, test items are written to the standards. At this time, decisions are made about which skills and knowledge can and should be measured, and test blueprints and item specifications that reflect these expectations are developed. After the assessments are developed, a post hoc review of alignment is conducted (La Marca, 2001).

Because tests consist of samples of items from a domain, whereas standards define the domain, it is almost impossible for any relatively short individual test to measure all aspects of the standards (Porter, 2002). This is particularly true if the test consists mainly of multiple-choice items. While multiple-choice items provide a breadth of content coverage that cannot be as easily achieved with open-ended formats alone (Hambleton et al., 2000), the multiple-choice format nonetheless limits the types of knowledge and skills that can be measured (Baker, O'Neil, and Linn, 1993; Bennett and Ward, 1993). This points to the need for an assessment system that includes multiple and complementary measures to provide greater coverage of the standards.

Aligning the QSAS to the Standards

The following represents an in-depth look at the steps that the SAO undertook to ensure that the QSAS, and particularly the QCEA, was fully aligned with the standards. Many of the stages overlap, so they should not be interpreted as strictly sequential.

Stage 1: Developing the Standards

In the summer of 2002, the Education Institute's CSO contracted with CfBT to develop curriculum standards in the four core subjects. CfBT delivered a set of draft standards to a panel of content experts in the spring of 2004 and revised the standards based on the comments it received. The final set of standards in math, science, and English was delivered in December 2004, and the final Arabic version of the Arabic standards was delivered in May 2005.

The standards are organized into categories, called *strands*, *headlines*, and *standards*. Strands represent the broadest level, while standards represent the smallest granular level (often referred to as *objectives* in other sets of standards). For example, in fourth-grade science, there are five strands: (1) scientific inquiry, (2) life science, (3) materials, (4) physical processes, and (5) earth and space. Under the life science strand, one of the seven headlines indicates that "students should be taught to know that animals produce offspring that become adults." Under this headline, there are three key standards: (1) describe the young of some common mammals; (2) recognize the young of some common animals other than mammals; and (3) recognize the main stages in the life histories of fish, amphibians, reptiles, birds, mammals, and insects.

Many standards are designated as "key standards" to signal to teachers that these are the most important standards to teach students. The other standards are meant to augment the key standards. In addition, the standards in all four subjects for 10th through 12th grades specify two sets of expectations, foundation and advanced.

Stage 2: Drafting the QCEA Items to Match the Standards

Between 2002 and 2004, the test developers worked from drafts of the curriculum standards that were being regularly updated. Because the deadline for the final versions of the standards was December 2004, yet operational logistics required the test developers to finalize their test items for the 2005 QCEA by November of 2004, it was impossible for the 2005 QCEA to be fully aligned to the standards. Particularly in the case of Arabic, the final version of the standards was markedly different from the draft versions provided to the test developers and, ultimately, was not finished until May 2005. ETS was contracted to develop

the English and Arabic tests, and CTB was contracted to develop the mathematics and science tests.

During the development phases of the standards, a number of meetings among the SAO, the RAND team, the test developers, and CfBT were scheduled to facilitate alignment of the test content to the standards to the maximum extent possible. In December 2003, the SAO organized a meeting in Doha for the test developers and CfBT to discuss the scope and sequence of the draft standards. This first meeting offered an overview of the structure and content of the standards so that the test developers could start thinking about developing items that would be appropriately aligned with the standards. A second meeting among CfBT, CTB, and ETS occurred in April 2004 in Doha. During this meeting, participants discussed how best to assess particular portions of the standards, reviewed example items that were aligned with the standards, and decided on future meeting dates.

After those initial meetings, a number of meetings were set up between the test developers and CfBT to review the draft standards in detail, brainstorm about the development of items that were aligned with the standards, review scoring rubrics for constructed-response items, and discuss which pieces of the standards were testable or not testable in a standardized manner, among other topics. In total, six meetings were held. In May 2004, CTB and CfBT met in Reading, UK, to discuss the mathematics and science standards for grades 4, 6, and 8. A subsequent meeting to discuss grades 5, 7, and 9 occurred in Monterey, California, in the United States, in September 2004. ETS and CfBT met in July 2004 in Princeton, New Jersey, in the United States, and in December 2004 in Amman, Jordan, to discuss the Arabic and English standards for all grades. The SAO invited members of the CSO to each meeting.

The various meetings allowed the test developers to discuss with CfBT their understanding of the testability of each standard; make decisions about which strands, headlines, and standards were testable in a standardized format; and prioritize among the strands, headlines, and standards, given the limitations of testing time and information and communication technology infrastructure. It was decided that the test forms would assess as many key standards as feasible.

The test developers also presented CfBT with example test items, and CfBT provided input and feedback on the items, noting which standards the items appeared to measure and whether those items partially or fully measured the standards to which they had been mapped.

Stage 3: Deciding Which Standards the QCEA Could Assess

After these meetings, the test developers provided the SAO with a brief report on which strands were considered testable in a standardized manner. They also provided an estimate of the number of items and length of testing time needed to reliably report results at the strand level. The test developers' reports suggested that between 120 to 160 minutes were needed to obtain meaningful results at the strand level (and even for that, some strands had to be combined with others because their percentage in the curriculum standards was not sufficient for separate reporting). They also indicated that a majority of the curriculum standards for each subject could be assessed using the current QCEA format, although a significant portion would also be better assessed outside of the QCEA.

For 10th through 12th grades, a decision was made to assess only the foundation level.

Stage 4: Auditing the Alignment of the QCEA to the Standards

In developing the QCEA, CTB and ETS outlined a set of test specifications, or blueprints, that delineated the number of items needed to sample the standards. These test specifications were then used to show how items or tasks on the test matched to those in the standards documents. The test specifications indicated that many of the standards were included on the QCEA.

While this procedure for demonstrating alignment may be useful as a preliminary guideline, it can also mask imbalances in the representation of the standards (Rothman et al., 2002). Thus, the SAO decided to examine alignment more formally through an external audit. In March 2005, the SAO issued a request for proposals to a selected group of contractors that had extensive experience providing advice and assistance to policy leaders on issues of test alignment to standards. The selected contractor, Norman Webb, held subject-specific alignment meetings over the summer of 2005, which members of the SAO attended. In a final report, Webb identified areas of alignment and areas where gaps exist between the QCEA tests and the standards in terms of content knowledge and performance expectations.

Table B.1 summarizes the report's findings. It is important to note that, in this first alignment study for both the Arabic and English tests, test developers and the alignment auditors had certain disagreements about the interpretation of the standards, the scope of the blueprint, and some methodological issues. Alignment criteria from the external audit are as follows:

1. *Categorical concurrence:* An important aspect of determining alignment between standards and assessments is whether both address the same content categories. The categorical-concurrence criterion provides a very general indication of alignment, if both documents incorporate the same content. The criterion of categorical concurrence between standards and assessment is met if the same or consistent categories of content appear in both documents. This criterion was judged by determining whether the assessment included items measuring content from each strand. The analysis assumed that the assessment had to have at least six items measuring content from a strand for an acceptable level of categorical concurrence to exist. The number of items—six—is based on estimating the number of items that could produce a reasonably reliable subscale for estimating students' mastery of content on that subscale.
2. *Depth-of-knowledge consistency:* Standards and assessments can be aligned not only on the category of content covered by each, but also on the basis of the complexity of knowledge required by each. Interpreting and assigning depth-of-knowledge levels to both objectives within standards and assessment items is an essential requirement of alignment analysis. Depth-of-knowledge consistency between standards and assessment indicates alignment if what is elicited from students on the assessment is as cognitively demanding as what students are expected to know and demonstrate as stated in the standards.
3. *Range-of-knowledge correspondence:* For standards and assessments to be aligned, the breadth of knowledge required by both should be comparable. The range-of-knowledge correspondence criterion is used to judge whether a comparable span of knowledge expected of students by a standard is the same as, or corresponds to, the span of knowledge that students need to correctly demonstrate in the assessment items or activities. The criterion for correspondence between span of knowledge for a standard and an assessment considers the number of key standards within the strand with one related assessment item or activity.

Table B.1
Summary of Alignment Audit for 2005 QCEA and Qatar Curriculum Standards

Audit Subject	Remarks
The standards in general	Reviewers noted that the Qatar curriculum standards compared favorably to those in Singapore and in California and Wisconsin in the United States. In both specificity and competency targets, the Qatar standards compared favorably with the other reviewed curricula. This was especially true for mathematics, where the standards had a somewhat higher level of expectations and more thoroughly covered the expected range than did the other reviewed curricula. For English as a foreign language, it was found that the required competencies were based on widely recognized international benchmarks. One major drawback was that the science standards were too numerous and too narrowly formulated.
Arabic	Alignment among the standards and the tests for grades 2–11 was found to need improvement. For grades 2–6, the main alignment issue was that the tests did not have sufficient breadth as represented in the standards. For grades 7–12, the main concern was that the tests measured an inadequate range of content for nearly all of the strands and with insufficient depth.
English	In general, the analyses indicated that the operational tests and the curriculum standards were partially aligned for all the grades. Full alignment was not achieved, mainly because too great a proportion of the items had a depth-of-knowledge level that was below the level of the corresponding standard and too small a proportion of the standards under some strands were not measured on the assessment.
Math	The analysis for grades 1–6 indicated that the assessments and the mathematics curriculum standards were partially aligned. This was due to the fact that the reasoning and problem-solving strand was not tested (it was considered more appropriate for classroom-based assessments) and partly due to the large number of standards under some strands. This made it very difficult to ensure that a minimum of one item mapped to at least half of the underlying standards, which is one of the alignment criteria. The results for grades 7–12 indicated that the standards and assessments were partially aligned. The lack of full alignment related primarily to two of the four alignment criteria, categorical concurrence and range-of-knowledge correspondence.
Science	Alignment with the key science curriculum standards was found to be good overall, with the exception of scientific inquiry, which was intentionally not tested because this strand was considered to be better assessed at the classroom level. Where there was a lack of full alignment between the science assessments and the standards, this was found to be due in part to a shortcoming of the science standards—the large number of fairly specific standards under each strand. The alignment report concluded that the Qatar science standards were more detailed and numerous than those in the comparison countries and states. The relatively narrow statements of expectations were not necessary and created a problem for alignment and with instruction.

4. *Balance of representation:* In addition to comparable depth and breadth of knowledge, aligned standards and assessments require that knowledge be distributed equally in both. The range-of-knowledge correspondence criterion considers only the number of key standards within a strand that have been addressed, or "hit" (a standard with a corresponding item); it does not consider how the hits (or assessment items/activities) are distributed among these key standards. The balance-of-representation criterion is used to indicate the degree to which one objective is given more emphasis on the assessment than another.

Stage 5: Developing Other Assessments Beyond the QCEA

As noted earlier, a single test, such as the QCEA, cannot assess all the skills and knowledge specified in the standards. As the QSAS evolves, other forms of standards-aligned assessments will supplement the QCEA to assess skills embodied in the standards, such as scientific inquiry skills, that are not appropriately assessed through a standardized paper-and-pencil assessment.

This includes performance-based tests that can measure applied skills or computer-delivered tests (in which the student takes the test on a computer, rather than on paper, but it is not adaptive).

APPENDIX C

Performance-Level Results of 2005 and 2006 QCEAs for Ministry of Education, Private Arabic, and Independent Schools

The tables in this appendix compare 2005 and 2006 QCEA results for students in Ministry of Education, private Arabic, and Independent schools in the four tested subjects: Arabic, English, mathematics, and science. Results for the tested grades (4–11) are presented in terms of performance levels: meets standards, approaches standards, and three levels demarcating that a student has performed below standards.

Table C.1
QCEA Performance Levels, Arabic, by School Type and Grade, 2005 and 2006 (percentage)

	School Type					
	Independent		Ministry		Private Arabic	
Grade and Performance Level	2005	2006	2005	2006	2005	2006
Grade 4						
Meets standards	6	3	3	2	3	1
Approaches standards	40	31	19	21	22	21
Below standards level 3	49	58	64	66	66	69
Below standards level 2	4	6	8	9	6	7
Below standards level 1	1	2	6	2	4	2
Grade 5						
Meets standards	8	7	3	3	4	3
Approaches standards	40	31	20	21	22	22
Below standards level 3	45	55	61	63	62	66
Below standards level 2	3	6	9	10	7	8
Below standards level 1	3	1	6	3	6	2
Grade 6						
Meets standards	9	7	3	4	5	4
Approaches standards	41	29	19	18	24	20
Below standards level 3	46	55	60	59	56	58
Below standards level 2	3	7	13	15	11	13
Below standards level 1	1	2	5	5	5	5
Grade 7						
Meets standards	11	6	4	4	3	4
Approaches standards	35	28	20	17	23	20
Below standards level 3	47	58	60	60	59	58
Below standards level 2	5	7	13	15	11	13
Below standards level 1	2	1	3	4	3	5
Grade 8						
Meets standards	12	4	4	3	5	3
Approaches standards	34	30	20	20	22	18
Below standards level 3	48	57	61	59	50	58
Below standards level 2	5	7	13	15	20	16
Below standards level 1	1	2	3	3	3	5

Table C.1—Continued

	School Type					
	Independent		Ministry		Private Arabic	
Grade and Performance Level	2005	2006	2005	2006	2005	2006
Grade 9						
Meets standards	9	4	4	3	5	2
Approaches standards	30	25	21	18	20	18
Below standards level 3	47	65	58	66	47	60
Below standards level 2	8	6	12	12	16	17
Below standards level 1	6	1	5	2	12	4
Grade 10						
Meets standards	20	7	3	2	5	2
Approaches standards	45	33	19	20	17	14
Below standards level 3	23	48	56	55	45	49
Below standards level 2	8	10	13	18	18	22
Below standards level 1	5	2	9	5	15	14
Grade 11						
Meets standards	19	9	3	4	3	3
Approaches standards	49	35	22	24	29	18
Below standards level 3	24	48	59	57	42	51
Below standards level 2	6	7	13	12	18	20
Below standards level 1	1	1	3	2	7	8

Table C.2
QCEA Performance Levels, English as a Foreign Language, by School Type, 2005 and 2006 (percentage)

	School Type					
	Independent		Ministry		Private Arabic	
Grade and Performance Level	2005	2006	2005	2006	2005	2006
Grade 4						
Meets standards	10	7	0	0	2	0
Approaches standards	32	21	3	3	9	4
Below standards level 3	46	48	50	49	55	56
Below standards level 2	8	9	30	17	25	16
Below standards level 1	3	15	17	31	10	24
Grade 5						
Meets standards	11	7	0	0	3	0
Approaches standards	27	24	5	5	13	7
Below standards level 3	46	45	49	46	51	50
Below standards level 2	8	9	16	18	11	16
Below standards level 1	8	14	30	31	21	26
Grade 6						
Meets standards	13	7	0	1	2	2
Approaches standards	32	21	5	7	8	9
Below standards level 3	41	52	51	51	51	53
Below standards level 2	4	8	9	15	8	16
Below standards level 1	10	12	35	26	31	20
Grade 7						
Meets standards	10	4	1	0	4	0
Approaches standards	22	15	6	5	12	10
Below standards level 3	44	57	53	53	51	49
Below standards level 2	10	10	16	16	13	17
Below standards level 1	14	13	24	25	20	24
Grade 8						
Meets standards	11	5	1	1	2	2
Approaches standards	27	14	6	5	12	8
Below standards level 3	40	58	51	59	48	58
Below standards level 2	10	12	16	16	12	17
Below standards level 1	13	11	26	19	26	15

Table C.2—Continued

	School Type					
	Independent		Ministry		Private Arabic	
Grade and Performance Level	2005	2006	2005	2006	2005	2006
Grade 9						
Meets standards	9	4	1	1	6	0
Approaches standards	27	11	7	5	11	7
Below standards level 3	36	46	50	40	39	42
Below standards level 2	16	25	22	31	19	27
Below standards level 1	13	14	21	23	24	23
Grade 10						
Meets standards	19	10	1	1	2	2
Approaches standards	34	20	7	5	10	8
Below standards level 3	30	43	50	46	43	35
Below standards level 2	8	14	21	24	21	27
Below standards level 1	9	13	21	24	23	28
Grade 11						
Meets standards	23	13	2	1	6	1
Approaches standards	39	24	7	8	10	9
Below standards level 3	25	40	49	49	46	44
Below standards level 2	3	15	21	24	16	28
Below standards level 1	9	8	22	18	22	19

Table C.3
QCEA Performance Levels, Math, by School Type, 2005 and 2006 (percentage)

	School Type					
	Independent		Ministry		Private Arabic	
Grade and Performance Level	2005	2006	2005	2006	2005	2006
Grade 4						
Meets standards	0	0	0	0	0	0
Approaches standards	33	41	17	17	19	16
Below standards level 3	37	42	42	43	45	47
Below standards level 2	17	12	18	20	16	20
Below standards level 1	14	6	23	20	20	18
Grade 5						
Meets standards	0	0	0	0	0	0
Approaches standards	39	37	14	18	23	22
Below standards level 3	39	44	40	44	41	48
Below standards level 2	13	13	23	25	20	20
Below standards level 1	9	5	24	14	16	11
Grade 6						
Meets standards	0	0	0	0	0	0
Approaches standards	55	48	18	17	24	20
Below standards level 3	23	39	31	44	27	39
Below standards level 2	7	11	17	23	16	25
Below standards level 1	14	3	34	17	33	16
Grade 7						
Meets standards	0	0	0	0	0	0
Approaches standards	48	49	20	18	22	17
Below standards level 3	28	34	31	42	28	36
Below standards level 2	8	13	16	24	13	27
Below standards level 1	15	4	33	15	37	19
Grade 8						
Meets standards	0	1	0	0	0	0
Approaches standards	51	49	17	15	19	16
Below standards level 3	23	35	34	48	33	45
Below standards level 2	6	11	11	24	11	19
Below standards level 1	20	4	38	13	37	19

Table C.3—Continued

	School Type					
	Independent		Ministry		Private Arabic	
Grade and Performance Level	2005	2006	2005	2006	2005	2006
Grade 9						
Meets standards	1	0	0	0	0	0
Approaches standards	45	50	18	19	19	13
Below standards level 3	23	27	32	38	28	32
Below standards level 2	15	10	21	17	16	16
Below standards level 1	17	13	28	26	37	39
Grade 10						
Meets standards	2	0	0	0	1	0
Approaches standards	47	61	23	19	19	19
Below standards level 3	15	11	26	26	18	17
Below standards level 2	9	2	16	7	14	2
Below standards level 1	27	25	35	48	48	62
Grade 11						
Meets standards	0	0	0	0	0	0
Approaches standards	69	48	23	25	29	28
Below standards level 3	15	15	27	14	26	13
Below standards level 2	1	0	3	0	2	0
Below standards level 1	15	37	47	61	43	58

Table C.4
QCEA Performance Levels, Science, by School Type, 2005 and 2006 (percentage)

	School Type					
	Independent		Ministry		Private Arabic	
Grade and Performance Level	2005	2006	2005	2006	2005	2006
Grade 4						
Meets standards	0	0	0	0	0	0
Approaches standards	41	16	17	14	16	13
Below standards level 3	42	42	43	47	47	49
Below standards level 2	12	20	20	20	20	20
Below standards level 1	6	21	20	20	18	17
Grade 5						
Meets standards	0	0	0	0	0	0
Approaches standards	37	25	18	24	22	25
Below standards level 3	44	40	44	46	48	49
Below standards level 2	13	20	25	21	20	18
Below standards level 1	5	14	14	9	11	8
Grade 6						
Meets standards	0	0	0	0	0	0
Approaches standards	48	31	17	12	20	14
Below standards level 3	39	41	44	41	39	41
Below standards level 2	11	19	23	30	25	29
Below standards level 1	3	9	17	17	16	17
Grade 7						
Meets standards	0	0	0	0	0	0
Approaches standards	49	18	18	15	17	14
Below standards level 3	34	36	42	43	36	31
Below standards level 2	13	30	24	31	27	34
Below standards level 1	4	15	15	12	19	20
Grade 8						
Meets standards	1	0	0	0	0	0
Approaches standards	49	35	15	14	16	11
Below standards level 3	35	41	48	46	45	42
Below standards level 2	11	17	24	28	19	28
Below standards level 1	4	8	13	12	19	20

Table C.4—Continued

	School Type					
	Independent		Ministry		Private Arabic	
Grade and Performance Level	2005	2006	2005	2006	2005	2006
Grade 9						
Meets standards	0	0	0	0	0	0
Approaches standards	50	30	19	11	13	11
Below standards level 3	27	37	38	39	32	26
Below standards level 2	10	15	17	25	16	21
Below standards level 1	13	18	26	25	39	42
Grade 10						
Meets standards	0	0	0	0	0	0
Approaches standards	61	39	19	16	19	7
Below standards level 3	11	25	26	44	17	30
Below standards level 2	2	15	7	17	2	21
Below standards level 1	25	22	48	24	62	42
Grade 11						
Meets standards	0	0	0	0	0	0
Approaches standards	48	43	25	22	28	25
Below standards level 3	15	27	14	35	13	31
Below standards level 2	0	8	0	14	0	13
Below standards level 1	37	22	61	29	58	30

References

Abedi, Jamal, "The No Child Left Behind Act and English Language Learners: Assessment and Accountability Issues," *Educational Researcher*, Vol. 33, No. 1, January–February 2004, pp. 4–14.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing*, Washington, D.C.: American Educational Research Association, 1999.

Ananda, Sri, *Rethinking Issues of Alignment Under No Child Left Behind*, San Francisco, Calif.: WestEd, 2003. As of January 30, 2009:
http://www.wested.org/cs/we/view/rs/693

Baker, Eva L., Harold F. O'Neil, Jr., and Robert L. Linn, "Policy and Validity Prospects for Performance-Based Assessment," *American Psychologist*, Vol. 48, No. 12, December 1993, pp. 1210–1218.

Bennett, Randy Elliot, and William C. Ward, eds., *Construction Versus Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment*, Hillsdale, N.J.: Lawrence Erlbaum Associates, 1993.

Brewer, Dominic J., Catherine H. Augustine, Gail L. Zellman, Gery Ryan, Charles A. Goldman, Cathleen Stasz, and Louay Constant, *Education for a New Era: Design and Implementation of K–12 Education Reform in Qatar*, Santa Monica, Calif.: RAND Corporation, MG-548-QATAR, 2007. As of January 14, 2009:
http://www.rand.org/pubs/monographs/MG548/

Broer, Markus, Juan E. Froemel, and Richard Schwarz, "Taught in a Foreign Language, Tested in the Mother Tongue: Advantage or Disadvantage for Test-Takers?" paper presented at the annual meeting of the American Education Research Association, Chicago, Ill., April 10, 2007.

Commission on Instructionally Supportive Assessment, *Building Tests to Support Instruction and Accountability: A Guide for Policymakers*, 2001.

Cronbach, Lee J., "Five Perspectives on Validity Argument," in Howard Wainer and Henry I. Braun, eds., *Test Validity*, Mahwah, N.J.: Lawrence Erlbaum Associates, 1988, pp. 3–15.

Gearhart, Maryl, and Joan L. Herman, "Portfolio Assessment: Whose Work Is It? Issues in the Use of Classroom Assignments for Accountability," *Educational Assessment*, Vol. 5, No. 1, 1998, pp. 41–55.

Hambleton, Ronald K., James Impara, William Mehrens, and Barbara S. Plake, *Psychometric Review of the Maryland School Performance Assessment Program (MSPAP)*, Baltimore, Md.: Maryland State Department of Education, 2000. As of January 30, 2009:
http://www.msde.state.md.us/Special_ReportsandData/Hambleton.pdf

Jolo, Hend, *Human Capital Formation in the State of Qatar with Special Reference to Oil and Gas Based Industries*, doctoral dissertation, Exeter, UK: University of Exeter, 2004.

Kane, Michael, "An Argument-Based Approach to Validity," *Psychological Bulletin*, Vol. 112, No. 3, November 1992, pp. 527–535.

———, "Validation," in Robert L. Brennan, ed., *Educational Measurement*, 4th ed., Westport, Conn.: American Council on Education/Praeger, 2006, pp. 17–64.

Koretz, Daniel, Daniel F. McCaffrey, Stephen P. Klein, Robert M. Bell, Brian M. Stecher, *The Reliability of Scores from the 1992 Vermont Portfolio Assessment Program*, Santa Monica, Calif.: RAND Corporation, DRU-159-EDU, 1992. As of January 30, 2009:
http://www.rand.org/pubs/drafts/DRU159/

La Marca, Paul M., "Alignment of Standards and Assessments as an Accountability Criterion," *Practical Assessment, Research, and Evaluation*, Vol. 7, No. 21, 2001. As of January 14, 2009:
http://pareonline.net/getvn.asp?v=7&n=21

La Marca, Paul M., Doris Redfield, and Phoebe C. Winter, *State Standards and State Assessment Systems: A Guide to Alignment*, Washington, D.C.: Council of Chief State School Officers, 2000.

Lane, Suzanne, "Validity Evidence for Assessments," paper presented at the Edward F. Reidy Interactive Lecture Series, National Center for the Improvement of Educational Assessment, Providence, R.I., October 14–15, 1999. As of January 30, 2009:
http://www.nciea.org/publications/ValidityEvidence_Lane99.pdf

Linn, Robert L., *Assessments and Accountability*, Los Angeles, Calif.: National Center for Research on Evaluation, Standards, and Student Testing, CSE Technical Report, November 1998.

Long, Vena M., and Christine Benson, "Re: Alignment," *Mathematics Teacher*, Vol. 91, No. 6, 1998, pp. 503–508.

Messick, Samuel, "Validity," in Robert Linn, ed., *Educational Measurement*, 3rd ed., New York: American Council on Education and MacMillan, 1989, pp. 13–103.

Pellegrino, James W., Naomi Chudowsky, and Robert Glaser, eds., *Knowing What Students Know: The Science and Design of Educational Assessment*, Washington, D.C.: National Academies Press, 2001.

Porter, Andrew C., "Measuring the Content of Instruction: Uses in Research and Practice," *Educational Researcher*, Vol. 31, No. 7, October 2002, pp. 3–14.

Qatar Ministry of Education, *Annual Statistics Report 2004/2005*, Doha, Qatar, 2005.

———, *Annual Statistics Report 2006/2007*, Doha, Qatar, 2007.

Qatar Planning Council, *Qatar Census 2004*, Doha, Qatar, 2005. As of January 14, 2009:
http://www.planning.gov.qa/Qatar-Census-2004/Flash/introduction.html

Qatar Supreme Education Council, "Admissions Policy," Web page, undated(a). As of January 30, 2009:
http://www.english.education.gov.qa/section/sec/education_institute/admission

———, "Qatar Comprehensive Educational Assessment (QCEA)," results for all years, Web page, undated(b). As of January 30, 2009:
http://www.english.education.gov.qa/section/sec/evaluation_institute/sao/_qcea

———, "School Performance Report Card," Web page, in Arabic, November 27, 2005. As of February 4, 2009:
http://www.education.gov.qa/SRC/search.htm

———, *QCEA 2005 Average National Scale Scores and Performance Level Results*, Doha, Qatar, 2006. As of January 14, 2009:
http://www.english.education.gov.qa/section/sec/evaluation_institute/assessment/_qcea2005

Roach, Andrew T., Stephen N. Elliott, and Norman L. Webb, *Alignment Analysis and Content Validity of the Wisconsin Alternate Assessment for Students with Disabilities*, Madison, Wis.: Wisconsin Center for Education Research, Working Paper No. 2003-2, 2003.

Rothman, Robert, Jean B. Slattery, Jennifer L. Vranek, and Lauren B. Resnick, *Benchmarking and Alignment of Standards and Testing*, Los Angeles, Calif.: Center of the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Technical Report 566, May 2002. As of January 30, 2009:
http://www.achieve.org/files/TR566.pdf

Shepard, Lorrie A., "Evaluating Test Validity," *Review of Educational Research*, Vol. 19, No. 1, January 1993, pp. 405–450.

Smith, Marshall S., and Jennifer O'Day, "Systemic School Reform," in Susan H. Fuhrman and Betty Malen, eds., *Politics of Curriculum Testing*, Politics of Education Association 1990 Yearbook, New York: Falmer Press, 1991, pp. 233–267.

Stapleman, Jan, *Standards-Based Accountability Systems*, Denver, Colo.: Mid-Continent Research for Education and Learning, April 2000.

Webb, Norman L., *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education*, Madison, Wis.: National Institute for Science Education, University of Wisconsin, Madison, Research Monograph No. 6, 1997.

———, *Alignment of Science and Mathematics Standards and Assessments in Four States*, Washington, D.C.: Council of Chief State School Officers, 1999.