THE HASHEMITE KINGDOM OF JORDAN
EDUCATION REFORM FOR KNOWLEDGE ECONOMY II (ERfKE II)

Mapping of Student Assessments in Jordan

Monitoring & Evaluation Partnership (MEP) Project

September 2014

This report is a product of World Education, Inc. (WEI) in collaboration with National Center for Human Resources Development (NCHRD) in Jordan under the Monitoring & Evaluation Partnership (MEP) project. MEP is a four-year (2010-2015) USAID-funded project implemented by WEI with the aim to strengthen the technical analytics capacity of NCHRD in M&E development and education policy studies and to provide financial support for a series of program quality evaluations for the Government of Jordan's Education Reform for Knowledge Economy (ERfKE II) program.

Acknowledgments

This study was conducted by the research team under the Monitoring and Evaluation Partnership (MEP) Program, which is supported by USAID in Jordan and implemented by World Education, Inc. in the U.S. and the National Center for Human Resource Development in Jordan. The authors wish to acknowledge the important contributions of Jordan's Ministry of Education (MOE) for their support and clarification of the assessment processes in Jordan. In particular, we are very thankful to many staff members from the Directorate of Examinations and Tests (DET), the select field directorates, schools, principals, teachers and students, and university professors and administrators for their openness and genuine interest in this important study. We would especially like to express our sincere gratitude to Nour Abu Al-Ragheb, USAID Education Specialist and Eng. Firyal Aqel, Director of the Donor Coordination Unit, for their unwavering support and coordination, without which this study would not have been completed successfully.

Acronyms

ANOVA	Analysis of Variance
CEA	Canadian Education Association
DET	Department of Examinations and Tests
EdData	Education Data for Decision Making
EGRA/EGMA	Early Grade Reading Assessment/Early Grade Math Assessment
ERfKE	Education Reform for Knowledge Economy
FGD	Focus group discussion
HS&B	High School and Beyond
IEA	International Association for the Evaluation of Educational Achievement
IRT	Item Response Theory
M&E	Monitoring and Evaluation
MCQ	Multiple Choice Questions
MEP	Monitoring and Evaluation Partnership project
MOE	Ministry of Education
NAfKE	National Assessment for the Knowledge Economy
NAfKE-JOR	National Assessment for the Knowledge Economy - Jordan
NCES	National Center for Education Statistics (United States)
NCHRD	National Center for Human Resource Development
NT	National Test (Jordan)
OECD	Organization for Economic Cooperation and Development (OECD)
PISA	Program for International Student Assessment
PIRLS	Progress in International Reading Literacy Study
SSME	School Management for Effectiveness
TIMSS	Trends in Mathematics and Science Study
UNRWA	United Nations Relief and Works Agency
WinDEM	Windows Data-entry Manager

Contents

Executive Summary

This study was conducted with support from USAID under the Monitoring and Evaluation Partnership (MEP) program, implemented by World Education, Inc.[1] in partnership with National Center for Human Resources Development (NCHRD) in Jordan. The study was requested by Jordan's Ministry of Education (MoE) to support the "Education Reform for Knowledge Economy" (ERfKE) program implemented by the MoE with international and domestic partners. The study objectives are to: 1) map out all major student assessments in Jordan and 2) identify elements for change, improvement and/or new development. The study has thoroughly reviewed and assessed the purposes—on paper and in practice—of the existing student assessments, the major domains of the core curricula subjects, data issues, including in comparison with international experiences, and other related assessment policy issues. The study also identified possible gaps and/or overlaps of the different student assessments in Jordan and suggested possible areas for policy consolidations or streamlining of certain assessments if necessary.

In Jordan, the major student assessments are:

1) The National Test (NT), a ***census-based*** test organized and administered by the MoE;

2) The National Assessment for the Knowledge Economy (NAfKE) test, a ***sample-based*** test organized and administered by National Center for Human Resource Development (NCHRD) created for the purposes of evaluating the ERfKE reform program;

3) The Tawjihii, the ***compulsory certification*** test for high school graduation exclusively organized and administered by a special unit in the MoE[2]; and

4) School Assessments, continuous or ***ongoing*** assessment which are carried out by teachers throughout school year but informed by guidelines from the MoE.

In addition, Jordan has been participating in the Trends in Mathematics and Science Study (TIMSS) assessment (since 1999), the Program for International Student Assessment (PISA) assessment (since 2006), and the Early Grade Reading Assessment/Early Grade Math Assessment (EGRA/EGMA) in which Jordan has participated are also included (since 2012).

As described in the report, we have "mapped" the current student assessment systems in Jordan, developed a new set of improvement recommendations

Methodology

This study has applied mixed methods to arrive at the intended results. We conducted a literature review to determine where Jordan stands in terms of quantity, "stakes" and utilization of student assessments in comparison to other countries. In addition, as the word, "mapping" implies, we reviewed and analyzed all the frameworks, purposes, assessment domains, grade levels, core subjects,

[1] World Education, Inc. is an international NGO, headquartered in Boston, USA. [www.worlded.org]

[2] We must note that Tawjihii will be mentioned in the report but requires and deserves a separate and significant analytical effort which is beyond the scope of this study.

test administration, guidelines, tools, relevant analytics, and reports of all major student assessment systems in Jordan, including the NT, Tawjihii, NAfKE, school assessment, and Jordan's participation in TIMSS, PISA, and EGRA/ EGMA. Focus group discussions (FGDs) with various education stakeholders, including teachers, students, school administrators, University of Jordan professors and administrators, and MoE staff in General Directorate of Examinations and Tests (DET), who developed and managed the student assessment systems were also carried out. Focus group participants answered both open-ended and close-ended questions. Furthermore, we gathered and examined a select set of key student assessment data sets from the multiple student assessment systems and compiled an integrated database.

Major Findings

- Among the diverse range of student assessment systems and models around the world, even from high-performing education systems that range from rigorous, high-stakes and frequent testing systems such as those in China, Korea, and Vietnam in Asia, to more flexible and low-stakes and infrequent testing systems such as those in Finland, the Netherlands, and Slovenia in Europe, Jordan, stands in the "middle" in terms of testing frequency, types of student assessments, grade levels, subject domains, as well as testing instruments, rubrics development, and administration. Jordan's student assessment systems as a whole are not by any means excessive or over burdening.
- The "stakes" of various types of student assessments in Jordan are sharply different from each other. For example, Tawjihii has the highest stake for students, teachers and schools and NAfKE has the lowest stake for the same stakeholders. This is not surprising in Jordan. In general, census-based or compulsory test tend to have high stakes for students and teachers and sample-based or optional test (including random selection of students) have low stakes. The table below summarizes the "stakes" findings from different education stakeholders' perceptions.

"Stakes" Attached to the Assessments in Jordan				
	Students	**Teachers**	**Schools**	**Avg.**
NT (by MoE)	3	4	5	4.67
NAfKE (by NCHRD)	1	1	1	1.00
School Assessment (MoE and all schools)	6	6	7	6.33
Tawjihii (MoE, special unit)	10	9	8	9.00
TIMSS (NCHRD)	1	2	2	1.67
PISA (NCHRD)	1	2	2	1.67
EGRA (MoE in partnership with USAID)	2	3	2	2.33
Explanatory Note: The ranked scale of 1 through 10 is given to each assessment system in Jordan based on qualitative data on "knowledge" and "perceptions" from focus groups with students, teachers, and school administrators and other educational stakeholders. 1-point means "I have never heard of it" or "not important to anyone". Some remarks often expressed as "I don't care and never prepare for it" or "never hear or think about it"; 10-point means "everyone knows it" or "everyone has to prepare for it". Some remarks expressed as "it is the most important test in my student career" or "my future depends on it" or "as a teacher, I have to teach to the test"				

MoE in Jordan may strategize to manage the level of "stakes" of the different kinds of student assessments to meet the purposes of the assessments as well as to make full use of the assessment

results for improving student learning and school quality. To take an action plan for improving school performance based on an assessment result is an example for raising the "stake"; to use NT assessment results as part of the continuous school assessment for grades 4, 8, and 10 is another example for raising the "stakes"; to encourage universities in Jordan to institutionalize multiple criteria for admitting students is an example for lowering the stakes of Tawjihii.

- Although there are multiple and diverse assessments in Jordan, there has been no or little attempt to integrate data from these diverse assessments. Nor has there been integrated analysis by linking multiple assessment results at student, class or school level to identify common problems and conduct higher-order analysis. As a result, cross-assessment validation, inter-assessment correlation, or between measurement-domain analysis could not be analyzed or studied.
- Although there are multiple and diverse assessments in Jordan, tracking individual students to examine their growth in learning remain impossible. As a result, school value-added evaluation or school effectiveness study with these assessment results is impossible. For example, the fixed interval of "every three years" under NT system to repeat a test in the same grade and frequent change of test items without an "item bank" make it impossible to track students over time and conduct school "value added" evaluation or school effectiveness evaluation.
- It became clear to us throughout the study that raw assessment data was by default not shared openly (e.g. no data is openly available on the MoE or NCHRD web site) unless there is "an official request" as one MoE officer explains. This is particularly true in the MoE. NCHRD shared some of its TIMSS and PISA assessment data with local universities in CD packets given the fact that international TIMSS and PISA data by countries is publicly available (downloadable from TIMSS or PISA websites). But in general, there has been limited data sharing. Many countries in North America, the European Union and developed nations in Asia start to make raw education data available online and downloadable for secondary data analysis and on-going use.[3] With the limited data sharing, one can only conclude that there has been no or limited secondary data analysis of student assessment results in Jordan. The MoE uses or analyzes NT data and NCHRD uses and analyzes NAfKE, TIMSS and PISA data. Perhaps the demand from other stakeholders is not present or perhaps there is a hesitation to make the raw data available by default for any further use. Whatever explanation might be, we believe that Jordan could start to champion the initiative of developing an open and transparent data system in education sector in the Middle East region.
- The utilization of all assessment data and results in Jordan has been assessed as insufficient in general. Two types of utilization are defined, use of data for analysis and use of the assessment results for policy planning or decisions. They are organically linked and should be positively correlated. Effective data analysis increases the likelihood of the results being used for policy actions or action plans. While data analysis is still within the theory of "information production", use of the analysis results for policy actions is considered as part of the "information consumption" theory. On the information production side, an integrated design of the student assessment systems is absent; integrated and higher order data analysis to address important policy inquiries is minimally enabled;

[3] We must note that in almost all cases in the world, the official downloadable data would conceal individual (student, teacher or school) identity. Raw data is only used for secondary data or statistical analysis.

presentation of the assessment results (reporting) are insufficient. On the information consumption side, there has been a lack of institutional and concerted efforts to facilitate action planning based on the newly produced information. Even if there is an action plan developed, there is a lack of follow up to monitor and evaluate the implementation of the new action plan. For NT report, 80% of the report each year simply covers descriptive tables or bar charts with little narrative, only reporting on averages or percentages of NT performance levels by school type, gender, and field directorate.[4] Although the NT annual report only targeted domestic educators (more likely for internal users with no English version), the annual report does not serve well to monitor student learning performance at the system level; does not identify real needy schools and students who may need more support or investment; and does not evaluate how satisfactorily students performed at national, regional and school levels. The simple averages presented fail to tell the "stories" of the NT performance levels, and the annual report therefore fails to capture the usefulness of the assessment data for informing policies.

Key Recommendations

- MoE should consolidate the current loosely-coupled student assessment systems into a well-integrated holistic student assessment system that includes various forms and designs of assessment tools for the purpose of improving equitable learning and achievement. Jordan should capitalize on the work already done in the area of student assessments so far and utilizes the existing local capacity, particularly within MoE and NCHRD to advance and upgrade the assessment systems with local and international technical assistance support. More specifically, three structural adjustments and integration are recommended in the integrated assessment system in Jordan: 1) Increase the level of effort to assess all students learning performance in Grades 4, 8, and 10 ***annually*** and ***raise the stakes*** of the NT assessment system. Institutionalize the development, delivery and dissemination of field directorate and school report cards and their utilization so that students and schools take NT "more seriously" and learn the academic subjects as they prepare for NT; 2) Convert NAfKE to NAfKE_JoR and expand its plan to assess samples of students in Grades 3, 6 and 9 making it a moderately high-stake test for evaluating the learning of the 21st century skills including critical thinking, problem solving, synthesis and communication skills, etc.; 3) continue to participate in TIMSS and PISA but add Grade 4 TIMSS, and institutionalize EGRA to benchmark an international "norm" of learning achievement. The next table (following page) illustrates the recommended structural adjustments.

[4] We fully understand that NT is census-based student assessment. When data is analyzed, there is no need to follow stringent inferential statistical norms or rules. However, certain type of descriptive statistical analysis from systems perspective is critical.

		Grades to be tested											
		1	2	3	4	5	6	7	8	9	10	11	12
Census Assessments	NT (annualized)				x				x		x		
	Tawjihi												x
Sample Assessments	NAfKE_JoR			x		x	x			x		x	
	TIMSS				x				x				
	PISA									15-yr. old			
Sample Assessment	EGRA		x	x									
MoE's ongoing assessment	School Assessment (SA)	All grades											

NT – National Test, managed and administered by the MoE, must have a sizable item bank of relevant domains, subjects, and grade levels. To raise NT stakes, results could be considered as part of school assessment for Grades 4, 8, and 10 students.

Tawjihi – General Secondary Certificate Examination in Jordan, currently managed and administered by the MoE

NAfKE_ JoR – National Assessment for Knowledge Economy Skills _Jordan, Reading and Math for grade 3 & reading, math and science for grades 5 and 9. It takes place every two years.

TIMSS – Trends in International Mathematics and Science Study (international). Jordan has participated in TIMSS for the past 4 cycles since 1999, but only in grade 8, not in grade 4. NCHRD manages and administers TIMSS

PISA – Program for International Student Assessment (international). Jordan has participated in PISA for the past 3 cycles since 2006. NCHRD manages and administers PISA

SA - School assessment is an on-going throughout an academic year, student's cumulated composite score may be from 1) subject learning performance (quizzes and tests), 2) discipline (behavior), 2) social responsibilities (peer support, community duties, and school tasks)

Note: Red color **X** indicates a new addition or major change, green color box means "cancelled", and dark color **X** means no change.

The new structure would let MOE plan student assessments with a long term vision. Not only will this new system have better integrated design and framework, but also it will enable more levels of higher-order analytics and more value-added benefits with the bigger and more useful data. The new system could potentially share the common-core item bank for test development and administration. It will make more "comparability" feasible across the time and administrative levels (central, district, school, class and student). For example, Grade 4 test results from two separate years are bound to be different, but how the MoE can decide to create anchor items in order to equate the two tests for the reliable comparability requires a strategic planning in the item design and item selection process. The education research community in Jordan—in tracking individual students over time about their learning outcomes as well as learning characteristics—may help explain the unexplainable phenomenon. For example, from this type of design, we may be able to answer why girls and boys in Jordan perform equally well in achievement results in early grades but very differently as they move up to higher grades. Boys underperform girls significantly in every subject and grade for the rest of the education career after Grade 4. Longitudinal comparison is not only an important statistical method but also a strategic thinking in terms of detecting and

identifying changes (progress or regress) over time and over "comparable elements." Only the new integrated assessment system would permit: a) tracking individuals' learning achievement over time and over similar criteria, b) tracking schools' performance over time and over different cohorts of students, c) tracking national trends over time and across similar performance measures. Analysts may also examine gaps or differences (in gender, among ethnic/migration groups, between rural and urban, etc.) and other variances between and within schools or directorates, as well as statistical relationships between and among outcome and explanatory variables. All of these require a good and strategic design and planning.

- All student assessments, particularly the new expanded NT system, must improve its analytics capacity to boost the potential utilization of the results. The analytics should at least include 1) longitudinal analysis (including tracking students and cohorts, 2) cross-unit analysis, and 3) comparative analysis against the national standards or expectations. New data analytics guideline and report template based on the new NT should be developed. The DET staff should be trained to conduct higher order data analysis. Co-partnership on this could be initiated but management and analytics should be quickly passed on. Since NT would be considered as the most important anchor system for other assessments within the new consolidated system for validation purpose and higher order research analysis, it must have ability to track individual students longitudinally with a smart design, which would allow MoE or education researchers to conduct the value-added analysis of the school effect.[5] USAID supported MEP project to strengthen NCHRD's analytics capacity is a successful project by all measures, and similar support should be replicated for the MoE and scaled up to the field directorate level.
- NT and NAfKE-JOR must have a sizable item bank for relevant domains, subjects and grade levels. Multiple categories of sub-domains could be envisioned for the bank and specific items under each subject or grade level could be developed and/or borrowed (or purchased) from credible and reliable sources internationally. For example, there are many items already developed to measure critical thinking skills, problem-solving skills, and synthesis skills in many test centers around the world for different grades and subjects. Jordan can surely match them to its own curriculum needs and the ERfKE requirement of the 21st century skills in addition to its own existing items embedded in the MOE and NCHRD. The Item bank development is known as an on-going development process which requires a significant national effort to manage, coordinate and maintain. Jordan should no longer wait and the MoE/DET and NCHRD may join forces to hire and train a few test item bankers.
- Teachers, supervisors and principals should improve data literacy skills. They must receive training to gain the ability to test students and interpret the results of those tests. Data literacy is a required competency for the improvement of classroom practices and students' academic performance. For academic performance, teachers must be at the forefront of this process. As recommended by Harvard educators who developed 8-step "Datawise" program for teachers, rapid assessment by and quick feedback to teachers is not only necessary for 21st century teaching and learning, but also

[5] The value-added school effect study requires tracking students and conducting several types of "learning gains" as the standardized outcome of the achievement and then singling out school net contribution (value-added) controlling for student, household and other social but outside school characteristics or factors.

a new required competency for any teaching profession[6]. If teachers have the ability to test students but do not have data literacy, testing will not be useful for teachers or students for an improvement purpose.

- MoE should raise the stakes of NT assessment. MoE could consider NT assessment results as a part of the school assessment for Grades 4, 8, and 10 students. For example, instead of having teachers develop their own final exams locally for these grades, NT assessment results could be used (up to 40% stake) in the final school assessment report. This will increase the perceived "stake" by students and teachers. It must be noted that the preparation for a test is a learning process, even in the current drive for the knowledge economy skill, particularly if the test items are measures of the critical thinking, problem solving and synthesis skills as the new curriculum promotes. We also strongly believe that Jordan's national report card, field directorate report card, and individual school report card, developed properly, would raise NT stakes. Revealing the NT performance results to all stakeholders through the comparative lens and the report card mechanism would contribute to greater transparency in developing the system-wide culture of data for educational decisions. It would surely bring about the higher stakes that the NT deserves.
- Although EGRA and EGMA has not been part of the overall MoE assessment yet, it is important for MoE to adopt it institutionally and use it as "early-stage detection" mechanism to identify specific learning needs, ineffective teaching and other related impediments to quality education and to support focused and remedial improvement programs timely. Eventually, ERfKE_JoR assessment for 3rd grade level could potentially substitute EGRA once it is fully developed and institutionalized.

Finally according to the highly publicized McKinsey's report (2007), "***How the world's best-performing school systems come out on top***", three factors that contribute to all high-performing education systems in the world are: 1) getting the right people to become teachers, 2) developing them into effective instructors and, 3) ensuring that the system is able to deliver the best possible instruction for every child. Clearly, none of the key determinants mentioned is about student assessment system development. The student assessment system alone won't be the determining factor for a high-performing education system, but a well-designed and administered student assessment system, with results used effectively, acting in unison with other smart education policies, should become essential for monitoring the system and individual performance levels and informing policy actions for the improvement of learning and teaching. Without it, the catchword, "improvement" is simply an empty verbiage.

[6] http://www.gse.harvard.edu/news-impact/2012/01/the-data-wise-process

I. Introduction

I.1. Background

This study was conducted with support from USAID under the Monitoring and Evaluation Partnership (MEP) program, implemented by World Education, Inc.[7] in partnership with National Center for Human Resources Development (NCHRD) in Jordan. The study was requested by Jordan's Ministry of Education (MoE) to support the "Education Reform for Knowledge Economy" (ERfKE) program implemented by the MoE with international and domestic partners. The Government of Jordan launched the second phase of the program, ERfKE II, in 2010.[8] The overall objective of ERfKE II is to provide students enrolled in pre-tertiary education schools (basic and secondary levels) with the knowledge and skills to participate in the 21st century knowledge economy.

To achieve this ambitious goal, ERfKE II focuses on five major aspects of education system development: 1) establishing and improving school-based management under the policy of decentralization, thereby empowering local school authorities and teachers to bring about the effective learning results; 2) establishing and enhancing institutional capacities of policy development, planning, monitoring and evaluation (M&E), and organizational management; 3) on-going review, development, and adjustment of teaching and learning resources with ERfKE II goals; 4) expanding and improving early childhood, vocational, and special education; and 5) improving education facilities and the overall schooling environment. The end goal is to improve student learning of the knowledge economy skills, such as improved problem-solving, analytical thinking, computer technology, and communications skills, as demonstrated in achievement results.

The study of "Mapping of Student Assessments in Jordan" (the Mapping Study) is part of a larger portfolio of external evaluation studies for the ERfKE II program. It cuts across all aspects of the reform agenda and addresses the overall mission goal of improving student learning of 21st century knowledge and skills. This study intends to better inform policy makers at the MoE and educational partners of the current student assessments in Jordan in terms of purposes, adequacy, "stakes" (how important the impact of testing results may have on students, teachers, and administrators), usability, and utilization strategies. It also aims to support new strategies for aligning the student assessments to the mission of teaching and learning the 21st century skills in Jordan. In this aspect, the Mapping Study is the most critical.

I.2. Objectives of the Study

The objectives of the Mapping Study are to: 1) map out all major student assessments in Jordan; and 2) identify elements for change, improvement and/or new development. More specifically, the study has thoroughly reviewed and assessed the purposes—on paper and in practice—of the existing student assessments, the major domains of the core curricula subjects, data issues, including in comparison with international experiences, and other related assessment policy issues. The study also intends to identify

[7] World Education, Inc. is an international NGO, headquartered in Boston, USA. [www.worlded.org]
[8] The 1st phase of the ERfKE program, called ERfKE I, was implemented from 2003 to 2009

possible gaps and/or overlaps of the different student assessments in Jordan and suggest possible areas for policy consolidations or streamlining of certain assessments if necessary.

As part of the study, we have developed a new set of strategies based on the results of this study for supporting the quality of the student assessments in Jordan, which may include but are not limited to purposes, types and features of the various assessments, mending gaps and reducing redundancies if any, process of test item development, test administration, and methods of analysis, process of routine reports and uses of the results for policy development and educational decisions.

We also recommend an integrated data system to enable higher-order assessment data analyses and utilization from multiple assessments and sources, multiple years and multiple levels. We strongly believe this as an important value-added process that is necessary—not only to maximize the utilization of student assessment data, but also to better inform education policy makers and stakeholders in Jordan of student performance trends over time and across field directorates as well as schools, so they can make better and more evidence-based policies and decisions.

I.3. Definition of Student Assessment

In this report, *student assessment* is broadly defined to include any student assessment in Jordan, including: national standardized tests or examinations and local school tests and assessments for summative or formative purposes; census-based or sample-based in design; compulsory or optional; developed by the MoE or externally developed by an outside agency such as NCHRD. The student assessment is carried out for the purpose of evaluating learning, determining streaming, promotion, certification, and/or accountability. In Jordan, the major student assessments include:

1) The National Test (NT), a census-based compulsory test organized and administered by the MoE for grades 4, 8, and 10;

2) The National Assessment for the Knowledge Economy (NAfKE) test, a sample-based optional test organized and administered by National Center for Human Resource Development (NCHRD) created for the purposes of evaluating the ERfKE reform program for grades 5, 9, and 11;

3) The Tawjihii, the compulsory certification test for high school graduation (grade 12) exclusively organized and administered by a special unit in the MoE[9]; and

4) School Assessments, which are informed by guidelines from the MoE but managed variably by individual schools for all grades.

In addition, for this report we will also discuss the Trends in Mathematics and Science Study (TIMSS) assessment, the Program for International Student Assessment (PISA) assessment, and the Early Grade Reading Assessment/Early Grade Math Assessment (EGRA/EGMA) in which Jordan has participated.

[9] We must note that Tawjihii will be mentioned in the report but requires and deserves a separate and significant analytical effort which is beyond the scope of this study.

The Mapping study is intended to sketch out all elements of the multiple student assessment systems in Jordan including the purposes, assessment domains, subjects, grades, frequencies, administration process, data analysis, and utilization.

II. Methodology

The Mapping study has applied mixed methods to arrive at the intended results. First, as the word, "mapping" implies, we reviewed and analyzed all the frameworks, purposes, assessment domains, grade levels, core subjects, test administration, guidelines, tools, relevant analytics, and reports of all major student assessment systems in Jordan, including the NT, Tawjihii, NAfKE, school assessment, and Jordan's participation in TIMSS, PISA, and EGRA/ EGMA. We also conducted focus group discussions (FGDs) with various education stakeholders, including teachers, students, school administrators, University of Jordan professors and administrators, and MoE staff who have developed and managed the student assessment systems. Focus group participants answered both open-ended and close-ended questions. In addition, we gathered and examined a select set of key student assessment data sets from the multiple student assessment systems and compiled an integrated database.

II.1. Study Questions

Because this study was requested by the MoE, a key feature of the study was to explore and identify the MoE's interests and questions in terms of the mapping of student assessments. For this, the research team met and collaborated closely with the relevant MoE staff and other stakeholders to jointly raise multiple research questions that would meet all parties' needs. Although the questions are a bit long—and some require further research steps beyond the current scope and level of this effort—we sought to include all inputs and suggestions from our partners. Our intention was to start the "ball rolling" through this study to further facilitate the discussion of student assessments in Jordan as well as advance the expectation of addressing the identified challenges in the near future.

Below are the research questions that helped determine the design and methodology of the study as well as the data collection and analysis process:

1) What do the varying student assessments of different levels (level 1: grades 1-4, level2: grades 5-10, level 3: grades 11-12) actually measure in terms of knowledge economy skills?

2) To what extent can national and international assessments be used to measure the trends in achievements over time?

3) To what extent do these domains (and sub-domains) overlap across the student assessment tools? Are they considered redundant or an unnecessary "burden"?

4) To what degree are these tests complementary, timely and relevant in terms of administrating dates, grades, and subjects?

5) How have the multiple student assessment data been analyzed, reported, and disseminated? If so, who were the beneficiaries of such information? Have any further actions been taken based on the shared information?

6) How have these test results been used in terms of informing and improving relevant programs at the MoE, directorate, school, teacher and student levels?

7) To what extent are these multiple test data comparatively analyzed so that some level of "validation or correlational" analysis can be carried out?

8) How do students assessments in Jordan differ by their goals and objectives? And what makes each of these assessments unique?

9) What are the gaps in the national and international student assessments that need to be addressed for them to be more informative about the quality of the education system in Jordan?

10) What would be the recommended plans/actions to make the most of those tests in order to enhance classroom practice and to improve quality of teaching and learning?

As one may find the scope of the questions a bit overwhelming, but we strongly believe that this is an important beginning step towards more widespread policy dialogue and debate on student assessments in Jordan—and how the MoE can best manage and benefit from them.

II.2. Literature Review and Document Reviews

For this study, we examined student assessment systems from multiple relevant countries in Asia, Europe and the Middle East. The objective of this examination was to anchor some references for comparative purposes, although the research team fully recognizes that each country must follow its own policy context and educational needs. Most importantly, we must reiterate that we have reviewed and relied on the purposes, frameworks, domain measures, data variables, and reports of all major student assessments in Jordan. This is the critical data that informed the Mapping Study.

II.3. Stakeholder interviews and Focus Groups Discussion (FGD)

Student, teachers, parents and supervisors participated in focus group discussions (FGD) that were conducted in twelve schools to examine perspectives on the student assessments that are systematically conducted in Jordan (e.g., NT, NAfKE, TIMSS, etc.). These discussion groups included male, female and co-educational schools, selected across the authority of the Ministry of Education, private schools, and United Nations Relief and Works Agency (UNRWA) schools. After the FGD, all stakeholders answered a 37-item questionnaire.

II.3.1 FGD Participant and Selection Process

One field directorate in each region of Jordan (north, middle, and south) was selected. Four schools within each field directorate were then selected, representing rural-urban, boys, girls, co-education classrooms, and large and small schools. To be eligible for selection, a school must have participated

in at least one of the following tests: NAfKE, TIMSS, PISA, or Tawjihi during the past few years. School performance levels on these tests were not part of the selection criteria. Teachers representing grades 4, 8, 10 and/or 12, and teaching Arabic, math, or science subjects were eligible for selection. Separate focus groups were held for male and female teachers. Student selection criteria included being in grade 8 or higher, and participation in at least two of the following assessments, NAfKE, TIMSS, PISA and/or Tawjihi. Parents of students in selected schools were also eligible to participate in the FGDs. Separate groups were conducted for male and female parent groups. Table 1 provides information on the number of FGD participants by school, region, school sex, and school authority. The participating schools are identified in terms of alphabet letters to protect the confidentiality of the FGD participants.

Table 1: Information on the number of FGD participants by school, region, school sex, and school authority

School	# FGD Participants			Region	School Sex	School Authority
	Teachers	Students	Parents			
A	9	15	7	North	Boys	MoE
B	6	15	7	North	Girls	MoE
C	10	15	8	South	Girls	MoE
D	14	16	9	South	Girls	MoE
E	5	9	5	South	Boys	MoE
F	12	15	12	Middle	Girls	MoE
G	10	16	6	Middle	Boys	MoE
H	10	14	1	Middle	Boys	UNRWA
I	6	8	13	North	Girls	UNRWA
J	7	15	11	Middle	Co-Ed	Private
K	7	8	6	North	Co-Ed	Private
L	10	13	9	South	Co-Ed	Private
Total	106	159	94			

II.3.2. Focus Group Discussions

FGD facilitators were trained to use a semi-structured discussion protocol. This protocol inquired about knowledge of the assessments conducted in Jordan, perceived quality of the assessments, consequences of assessment in general, student attitudes about assessment, and so forth (FGD protocol is available upon request). The FGD protocols were similar across all respondent groups (teachers, parents and student), but varied appropriately depending on the participant group. Initial questions were systematically asked of the FGD participants. Facilitators probed for more specific detail when needed or to stimulate further discussion. FGDs were audio taped and summary notes were taken. The summary notes were reported in both Arabic and in English. Time and resource constraints required the use of the summary notes as the unit of analysis rather than audiotape transcriptions. Audiotapes were used when the summary notes were not clear and as a check for accuracy from time to time. The Arabic notes were more comprehensive in terms of descriptive details than were the English translated notes. NCHRD researchers used the Arabic notes in developing the coding structure and in the qualitative analysis of the FGD data as these were more detailed.

To develop the coding structure and establish inter-rater reliability of developed codes, both the Arabic and English summary notes were used. NCHRD researchers reviewed both Arabic and English summary notes. An external evaluator reviewed the English translation of the summary notes to establish inter-rater reliability of both the development of the qualitative data coding system and the analysis of the FGD data (see below).

II.3.3. Development of the Coding Structure

The primary codes were informed by the FGD protocol. Using the Arabic and English summary notes, NCHRD and the external evaluator independently identified primary coding areas and sub-codes under each of the primary code areas. Results were shared. Perfect agreement was reached in terms of coding areas and definition of those areas for all but one code area definition. Using a consensus process, disagreement was easily resolved, and agreement on the coding system was reached.

II.3.4. Characteristics of the Questionnaire

The 37 items in the questionnaire were grouped into two themes: 1) Level of importance of each assessment (rated as 1= the most important 2=somewhat important, 3= the least important); and 2) Level of usefulness of each assessment (rated as 1= the most useful; 2=somewhat useful; 3= the least useful). The results from the tabulation of those items were used to support the FGD notes. In addition to the items, questionnaires also included two open questions to ascertain: (a) participants' opinions about the three most important reasons why female students outperform male students academically; and (b) participants' opinions about the main reasons for the overall decline in students' academic achievement (males and females) in the last few years.

II.4. Gathering Student Assessment Data in Jordan

We first tested if we were able to collect and organize the historical data on all major student assessments in Jordan (such as NT, Tawjihii, TIMSS, PISA, NAfKE, and School Assessment) to see if the data could be obtained, integrated and used for a value-added purpose. With supports from the main stakeholders, the MoE and NCHRD, we were able to collect and organize limited but multiple years of data from all major student assessments in Jordan. The MoE provided three years of NT data[10], (Grade 8 in 2007, Grade 10 in 2011, and Grade 4 in 2012) and three years of Tawjihii data (2010, 2011, 2012). NCHRD provided all cycles of TIMSS data (1999, 2003, 2007, and 2011)[11], PISA (2006, 2009, and 2012 for 15 year olds), and NAfKE (2006, 2008 and 2011, all for Grades 5, 9, and 11). The process of organizing these large datasets itself was significant in that it began answering questions such as: 1) Are all the student assessment data available and obtainable? 2) Is it possible to integrate them all and at what level can that be done? 3) How likely is it that any higher order data analysis could be conducted with the existing data? 4) What could potentially be done (value-added) in the future to benefit from the existing assessment results?

[10] MoE has more than 10 years of NT data. However, according to the officials, due to various major changes in student assessment systems, year over year comparative analysis over a long time is not possible. In addition, MoE only manages and administers one grade each year so each of Grades 4, 8 and 10 gets cycled every third year.
[11] Note: Only Grade 8 students participated in TIMSS in Jordan. Grade 4 students never participated in TIMSS in Jordan.

II.5. Limitations

All key stakeholders in Jordan have strong interest in increasing knowledge about all aspects of student assessments. Much money and efforts have been spent on the diverse assessment tools for various purposes, key academic subjects and different grade levels, but many questions remain unanswered about the benefits and usefulness of these assessments for improving the quality of education in Jordan. While this study is an initial step towards better understanding a wide range of aspects of student assessments in Jordan, we must recognize some limitations in terms of the scope of this study. First, this study will not provide an in-depth analysis of the Tawjihii, the secondary school certification test that also qualifies pre-tertiary graduates for higher education in Jordan. Second, this study will not analyze test item pools within each assessment instrument, such as how reliable or valid the items are in measuring the test constructs or domains and how a difficulty index could be developed or referenced. Thirdly, our initial attempt to identify cost related information on various assessments proves in vain. Given the most domestic tests or school assessment are developed and managed by the MoE, it is largely person-month work and level of efforts that could not be possibly translated to cost information. Although the cost associated with data collection for sample-based test such as NAfKE, TIMSS and PISA could be averaged at about 40,000 US dollars, the cost associated with test development, data analysis, training, report writing, etc. could hardly be obtained, and vary tremendously from one assessment to another. Therefore, this study does not include the cost information. Finally, while this report will demonstrate select examples of analysis that could be carried out with the test data in the future, either test alone and integrated with other data, the study will not conduct data analysis for each test data set or present assessment results for policy actions or plans.

III. Literature and Synthesis

The world-wide literature review reveals that student assessment systems often require a good system design and development, but more importantly, demand a strong policy attention and ongoing support from institutions. Managing various "features" of the student assessment system is no easy task. The Canadian Education Association (CEA) recently described various characteristics of student assessments in an article entitled "Standards, Accountability, and Student Assessment Systems:"

- Low versus high stakes for students and schools (teachers and principals/school administrators);
- Internally versus externally developed or administered;
- Nationally versus regionally oriented;
- Geared toward all ages of school children versus key developmental points;
- Geared toward a variety of subject areas or a select few core subjects;
- Geared toward academic versus non-academic domains;
- Traditional paper-based modes versus technology-enhanced delivery modes;
- Reported at the student, school, and/or district level and national level;
- Focused on assessment *of* learning versus assessment *for* learning.

Many education systems have blended mixtures of these features in the development of their student assessment systems to serve various demands and purposes. These assessments are often offered in different forms, at different time intervals, for various subjects and administered with different grades or age groups. Jordan is of no exception. However, as mentioned earlier, the quality of these assessment tools varies tremendously and the results of these assessments may often lead to very different educational decisions or action plans for different stakeholders. For example, large scale national assessments often serve the purpose of improving school learning for education policy makers, while other standardized aptitude tests are typically used as filtering mechanism for admission to higher level of learning institutions, local or classroom assessments are more likely to serve for promoting and repeating students, and/or others may serve to check against pre-agreement of accountability (MOE in Singapore, 2011; CCSSO, USA; 2010; NAEP, USA, 2013; CIEB, 2012).[12]

III.1. European Systems

To get a glimpse of the student assessment systems of many countries in Europe, we adopted the following table(2), from the European Commission to show information about all national tests of student learning performance in each of the countries (Eurydice, EC 2009).[13]

Table(2): Information about all national tests of student learning performance in each of the countries.

Number and type of national tests in Europe, ISCED levels 1 (primary education) and 2 (lower secondary education), 2008/09				
Country	**Compulsory tests**	**Sample tests**	**Optional tests**	**School years in which they are administered**
Belgium – French Community	1			Year 6 of primary education
Belgium – Flemish Community		2		Years 6 and 8
Bulgaria	3			Years 4, 5 and 6
Denmark	10			Between years 2 and 8
	1			Year 9
Germany	1			Year 9
Estonia		2		Years 3 and 6
	1			Year 9
Ireland	3			End of 1st class/beginning of 2nd

[12] CCSSO is an acronym for Council of Chief State School Officers (http://www.ccsso.org/); NAEP is National Assessment of Educational Progress, which is the largest nationally representative and continuing assessment (USA) of what America's students know and can do in various subject areas, http://nces.ed.gov/nationsreportcard; CIEB stands for Center on International Education Benchmarking (http://www.ncee.org/programs-affiliates/center-on-international-education-benchmarking/)

[13] Source: National Testing of Pupils in Europe: Objectives, Organization and Use of Results, Education, Audiovisual and Culture Executive Agency, 2009 pages 27 and 28 (Figure 2.2).

				class; end of 4th class or beginning of 5th class; year 3 of post primary education
		2		Second and sixth classes
Spain		1		Year 4 of primary education
France		4		Two at the end of primary education, and two at the end of compulsory education
			1	Year 3 of primary education (known as 'CE2')
	1			First year of lower secondary education
Italy		3		Two in year 5 of primary education; one in year 1 of lower secondary education
	1			Year 3 of lower secondary education
Cyprus	1			Year 6
Latvia	3			Years 3, 6 and 9
Lithuania		2		Every year in either years 4 and 8, or years 6 and 10
			1	Year 10
Luxembourg	3			Years 3 and 6 of primary education, and year 5 of secondary education
Hungary	3			Years 4, 6 and 8
Malta	8			Years 4 ,5 and 6 of primary education; 1 to 5 of secondary education
			2	Year 6 of primary education; end of secondary education
Netherlands			1	Final year of primary education
Austria		2		Years 4 and 8
Poland	2			Year 6 of primary education; Year 3 of lower secondary education
Portugal	3			Years 4, 6 and 9
Romania		1		Year 4
	2			Years 7 and 8
Slovenia			1	Year 6
	1			Year 9
Slovakia	1			Year 9
Finland		2		Years 6 and 9
Sweden	3			Years 3, 5 and 9
England	2			Years 2 and 6
			5	Years 3, 4, 5, 7 and 8
Northern Ireland			1	Year 6

Scotland			6	Five tests in the National 5-14 Assessment Bank and one test in year 4 of secondary education
		4		Years 3, 5 and 7 of primary education; year 2 of secondary education
Iceland	2			Years 4 and 7
Norway	4			Years 2, 5, 8 and 10

Explanatory notes:

5 countries or regions within Europe do not have national tests: 1) Belgium German speaking region; 2) Greece; 3) Czech Republic; 4) Wales; and 5) Liechtenstein.

Ireland: The sample tests are carried out every five years.

Spain: The test in year 4 of primary education is taking place for the first time in the 2008/09 school year. A second test in year 2 of lower secondary education is being administered for the first time in the 2009/10 school year. These two tests will take place every three years.

Lithuania: In the sample tests, school years 4 and 8 are tested in odd (calendar) years, and school years 6 and 10 are tested in even years.

Netherlands: While participation in the test is at the discretion of the school or authority concerned, in practice nearly all pupils take it.

Finland: In most cases, one or two sample tests are administered each year. Generally the tests take place in school years 6 and 9, or at other curricular transition points.

United Kingdom (ENG): Tests at the end of year 2 are used to support the teacher assessment process and not reported on separately. Optional tests are used by the majority of schools but they are not statutory.

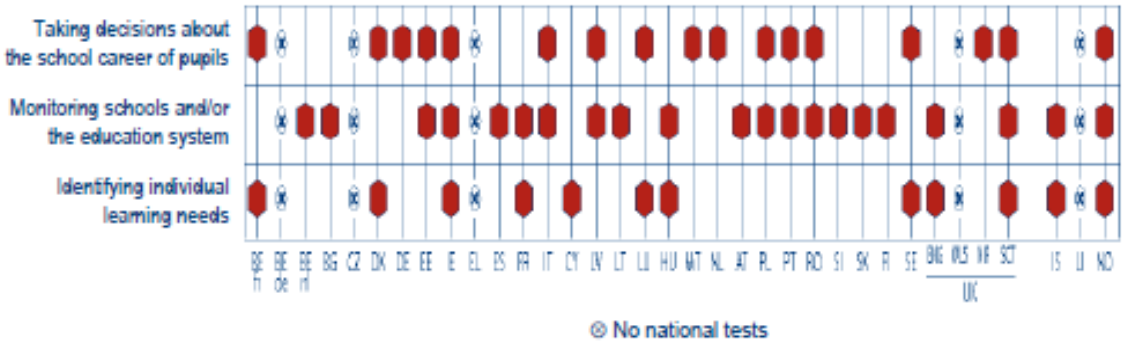
United Kingdom (SCT): During nine years of education and depending on their progress in class work, most pupils take five tests from the National 5-14 Assessment Bank. Though these tests and the national examinations in year 4 of secondary education are in principle optional, almost all pupils take them.

Iceland: The nationally coordinated examinations in school year 10 will resume from 2009/10.

As it is recognized, European education systems have a wide range of diverse student assessment systems. Not only are they testing in different grades, subjects, scales, but they are also different in their objectives. The table below (Eurydice, European Commission 2009) shows the different purposes of their national tests in different countries.[14] As one may see there are three major objectives of the national tests: 1) taking decisions about the school career of pupils (promotion, streaming, or tracking), 2) monitoring schools and/or the education system (identifying problems, evaluation for actions), and 3) identifying individual learning needs. Evidently, many European countries use the national standardized tests for one, two or all three objectives. The mix varies clearly.

[14] **Source**: National Testing of Pupils in Europe: Objectives, Organization and Use of Results, Education, Audiovisual and Culture Executive Agency, 2009, page 25 (Figure 2.1).

Main aims of nationally standardized tests, ISCED levels 1 and 2, 2008/09

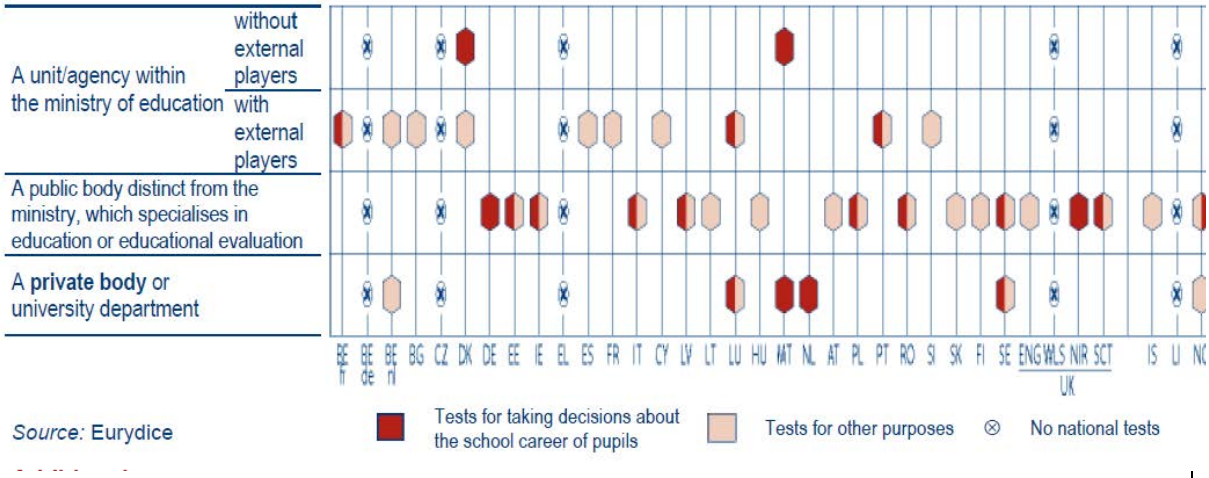


Explanatory note:
Only the main aim of each national test is represented in the Figure. Countries are allocated more than one category if they administer several tests with different main objectives.
France: A written examination with content standardized at national level is organized in several subjects for the award of the national certificate (the brevet) at the end of lower secondary education. Despite the existence of central procedures for administering and marking this examination, it cannot be regarded as a form of nationally standardized testing, given the wide variety of practices adopted in marking and interpreting its results.
Poland and Iceland: One or more national tests have two equally important objectives.

A great majority of the countries in Europe use the national standardized test for monitoring school and education system performance as indicated in the middle band of the chart above. For those which do not use the national standardized test for identifying individual learning needs, there have often been classroom assessments (non-standardized) for identifying individual learning needs. This is quite similar to the current case in Jordan. While NT serves the purpose of monitoring the national and directorate-level performance, the on-going classroom assessment (school assessment or student portfolio assessment) in Jordan is intended to identify the learning needs. More will be discussed later.

In terms of internal and external responsibilities of designing and managing national tests, there is also quite a diversity of arrangements within Europe. The following table (Eurydice, European Commission, 2009) shows the mixed arrangements. It is evident that most European countries with national tests have these tests either outsourced to a public body or agency rather than the MOE or have the MOE work in collaboration with an external agency for various testing purposes.

Bodies responsible for setting national tests, ISCED levels 1 and 2, 2008/09



Source: National Testing of Pupils in Europe: Objectives, Organization and Use of Results, Education, Audiovisual and Culture Executive Agency, 2009, page 42 (Figure 2.7).
Malta: The Matriculation and Secondary Education Certificate Examinations Board at the University of Malta sets the tests for the secondary education certificate examination taken at the end of compulsory education. The other tests are compiled by the directorate for quality and standards in education within the Ministry of Education.
United Kingdom (NIR): The last centrally provided transfer tests were taken in November 2008 for entry in September 2009.

A common practice across Europe in administering national tests is to ask the current teachers[15] to administer a given national test. They are often trained on specific marking instructions and guidelines specific to each test. Although many teachers who administer the tests are from the same schools as students, they do not teach the same students. Regional or local educational authorities would make unannounced visits to targeted schools during the time tests are administered.

In terms of the utilization of the student assessment results in Europe, we learned that the most of the national assessments have low stakes for students but "have important roles in policy planning and actions for policy makers. Often the assessment results are analyzed when formulating measures to deal with disparities, achievement gaps, and informing ongoing professional development for teachers" (Eurydice, 2009). In addition, many countries provide schools with their aggregated test results for comparisons with the national averages and subnational averages. However, schools are left to decide how they will use these results to improve their work if they are not satisfied with their own performance. According to European Commission's report by Education, Audiovisual and Culture Executive Agency, "unlike the United States and Canada, the testing results in many of the European countries are rarely used as an accountability tool which involves sanctions and rewards, and may affect resource allocation."

We must note that data on student assessments results is quite available and accessible in many countries in North America and Europe for secondary data analysis. For example, in the United

[15] Even when a national test is outsourced to other agencies, they still use school teachers to administer the test.

States, raw student assessment data and any other teacher and school data are all available online for any types of secondary data analysis by researchers and university students. According to the National Center for Education Statistics in the United States (NCES), one data set, called "HS&B" (High School and Beyond) featuring a longitudinal tracking study, has been used by hundreds of master and doctoral candidates in many universities in the United States for their theses and dissertations.[16] This is indeed value-added by making data available for secondary data analysis.

III.2. Select Systems in the Middle East

Several countries in the Middle East are considered the most relevant comparable countries to Jordanian context. Lebanon, Tunisia, Bahrain, and United Arab Emirates are similar to Jordan in history, size, and education systems. In addition, all four countries participated in TIMSS and three of them will participate in PISA 2015, and Tunisia and Lebanon, in particular, are similar to Jordan in their resource constraints but their educational systems are considered to be successful relative to other Arab countries. Their student assessment systems, then, may provide useful and relevant information for the readers of this report. The following section provides some more specific relevant information regarding the national student assessment (http://timessandpirls.bc.edu).

Lebanon

Lebanon's education system has two types of student assessments: school and central examination systems. The school examination system, which is locally designed and developed for all students attending public schools in Cycles 2 and 3 (equivalent to Grades 4 – 9, and secondary level students, Grades 10—12). In this system, two examinations (mid-term and end-term) are administered throughout the school year. Additionally, monthly tests and quizzes which are a part of the overall school examination system are conducted. Private schools apply the same school examination system as public schools but they take three term examinations or tests each school year. Although schools develop assessment tools variably, the central MoE provides guidelines for schools, similar to the system used in Jordan. Details of the guidelines were not available for this study.

The "central examination" system in Lebanon is meant for all students in public or private schools, but not all grades. Students are required to take the central official examination at the end of the basic education stage (Grade 9) to obtain an intermediate certificate. Those who pass may be eligible to pursue the secondary education. At the end of the secondary stage (Grade 12), students are required to take another one which consists of four major subjects, general science, life science, economics and sociology, and arts-humanities. Students who pass the central official examination

[16] The HS&B survey included two cohorts: the 1980 senior class, and the 1980 sophomore class. Both cohorts were surveyed every two years through 1986, and the 1980 sophomore class was also surveyed again in 1992. Other similar studies, longitudinally tracking students include. NELS:88 started with the cohort of students who were in the eighth grade in 1988, and these students have been surveyed through 2000; ELS:2002 began with a cohort of high school sophomores in 2002. This cohort will be followed through 2012; HSLS:09 began with a cohort of ninth graders in 2009. The first follow-up is planned for 2012 when most of the students will be high school juniors. More details could be found "https://nces.ed.gov/surveys/hsb/"

will obtain a General Secondary School Certificate and may be eligible to enter university. This is very much like Tawjihii in Jordan or Thanaweya Amma in Egypt.

In Lebanon, school assessment results are the determining factor for student promotion or repetition from 3rd Cycle onwards since students in the first and second cycles do not repeat regardless of performance. The central official examination is used to evaluate the preparedness of students for graduation and admission to a higher level of educational pursuit. However, one may wonder how students are supported if they do not do well in the school examinations in earlier grades? It appears that there is a lack of standardized student assessment for ongoing monitoring of school quality in earlier grades in Lebanon.

Tunisia

The educational system in Tunisia conducts several optional examinations at different stages. At the end of Grade 6, students may take an optional examination. Those who perform exceptionally well on it are selected to continue their education in special lower-secondary schools for the gifted or outstanding students. Similarly, at the end of Grades 9, students may again take another optional examination and those who perform exceptionally well will continue in special upper-secondary schools for the gifted or outstanding students. Those who choose not to take the exam or performed poorly remain in regular schools. At the end of upper-secondary education (Grade 13), students take the National Baccalaureate Examination (*Examen National du Baccalaureate*), the content of which consists of six subjects. Each examination subject is assigned a weight depending on the student's course of study, and the average of these weights determines the student's exam grade. Students who pass the baccalaureate can enter the university, while those who do not pass enter the workforce or study at a private school.

Bahrain

In 2007, the Evaluation and Assessment Center within the MoE in Bahrain introduced a new school evaluation system in all schools in Bahrain in order to accurately monitor student progress in all subjects. The new evaluation system assesses student performance, attitudes, and behaviors through daily class work, homework completion, classroom quizzes and tests, and a final assessment result which consists of teachers observations of student behavior, attitude toward classmates, and subject knowledge. This final assessment takes up to 30 percent of the student's final mark, while the midterm examination is worth 20 percent. Project work (by both individual and group), which varies across subjects, accounts for another 20 percent and the end of term examination is the final 30 percent. None of the examinations are standardized.

However, as legislated in Article 4 of the Royal Decree, the Quality Assurance Authority for Education and Training is mandated to review the quality of the performance of education. Within the Quality Assurance Authority, the National Examination Unit is responsible for evaluating student

learning progress at Grades 3, 6, and 9 in the four major subjects: mathematics, science, Arabic, and English. The unit conducts standardized examinations in these four subjects and collects information about the students, schools, and school performance. In May 2009, the National Examination Unit implemented examinations in all schools at Grade 3 (Arabic and mathematics) and Grade 6 (Arabic, mathematics, English, and science). In 2010, the examination was implemented at Grade 3 (Arabic and mathematics), Grade 6 (Arabic, mathematics, English, and science) and Grade 9 (Arabic, mathematics, English, and science), again for all public schools. This is the largest and most comprehensive national assessment in Bahrain to date with the aim of improving education quality and student learning performance. So far, however, the results of the National Examination Unit's examinations have not been made public.

United Arab Emirates

Continuous assessment occurs in all grades in public schools in the United Arab Emirates (UAE). The purpose of this assessment is to monitor student learning progress in key subjects, although the assessment is not standardized. For the school level continuous assessment, different evaluation tools are used, depending on the grade and the subject. For example, students in Grades 1–5 are assessed with written tests prepared by their teachers at the end of each textbook unit in each subject. Other evaluation tools include the following: classroom activities (such as oral presentations, written activities, and practical exercises) and non-classroom activities (such as research projects and portfolio development). According to the MoE, students are promoted to the next grade automatically during the primary education level. However, if a student does not achieve 50 percent on the total examination score, he or she will be enrolled in a remedial program at the end of the school year. If a student still fails, he or she will enroll in another remedial program at the beginning of the following year to support his or her learning in the next grade.

Students in Grades 6—9 are assessed using the similar assessments in place for students in Grades 1–5 in both mathematics and science. These students also take short written tests. Students need a total score of 50 percent in each subject to pass and be promoted to the next grade. However, if a student fails an examination in any given subject (up to a maximum of three subjects), he or she is allowed to retake the exam at the end of the academic year and before the summer holiday. If the student fails the exam again, he or she must repeat the grade.

At the end of each semester, public school students in Grades 1—9 receive a report card, which includes the scores obtained in each subject and level of evaluation, as well as any promotional comments or observations related to remedial programs from teachers of all subjects. Teachers record the standards of student performance and areas of improvement, which are presented to students' parents or guardians periodically with teacher recommendations, notes, and evaluations.

III.3. Summary

An international literature review of student assessment systems informs that there is a wide and diverse range of student assessment systems and models around the world. These diverse assessment systems are characterized by many different purposes, domain coverage, measures of knowledge and skills, and administering procedures. Although each country makes its own decisions about what student assessment systems should be, all "shapes and forms" of student assessments can be found in the world. While the quality and quantity of student assessments is almost impossible for comparative analysis for lack of detail or in-depth analysis in literature, high-performing education systems are demonstrated in both ends of the scale, from rigorous, high-stakes and frequent testing systems such as those in China, Korea, and Vietnam in Asia to more flexible and low-stakes and infrequent testing systems such as those in Finland, the Netherlands, and Slovenia in Europe. It is evident that there has been no fixed reference or "best model" to help answer the question "to what extent or how often should students be assessed or tested in Jordan or in today's education context?"

Comparatively, we conclude that Jordan, given the current student assessment systems, stands in the middle of the diverse student assessment systems in the world. In terms of testing frequency, types of student assessments, grade levels, subject domains, as well as testing instruments, rubrics development, and administration, Jordan's student assessment systems as a whole are not by any means excessive or over burdening. The school "testing culture" and "stakes" of student assessments in Jordan are moderate. However, the utilization of the assessment results for meeting the intended purposes is inadequate by our analysis and evaluation. More detailed findings and argument will be presented later in the report.

IV. Mapping of Student Assessments in Jordan

To map out student assessments, the study has analyzed each student assessment system in Jordan and listed various purposes, domain measures, administration processes, and utilization. We have also reported on how education stakeholders think of these assessments in terms of necessity, importance, and usefulness.

IV.1. Current Student Assessments in Jordan

As described earlier, 4 major domestic student assessments and 3 international student assessments are in Jordan: 1) National Tests (NT) for Grade 4, 8, and 10 in four core subjects, Arabic, Math, Science and English; 2) National Assessment for Knowledge Economy (NAfKE) for Grades 5, 9, and 11 in Arabic, Math and Science subjects; 3) Tawjihii, the Grade 12 graduation certification test in all learned subjects; 4) Local School Assessment (LSA) for all grades and on all core subjects. In addition, Jordan has been participating in two international tests: 5) Trends in Mathematics and Science Study (TIMSS) for Grade 8[17] in Math and Science subjects since 1999; 6) Program for International Student Assessment (PISA) for 15 year olds in Reading, Math and Science subjects since 2006; and 7) Jordan

[17] TIMSS assesses both Grades 4 and 8 students worldwide, but Jordan only participates in 8th grade student assessment.

also recently began to participate in Early Grade Reading and Math Assessments (EGRA and EGMA)[18] for Grades 2 and 3 students. The following table (3) shows an overall map of the various tests in Jordan by levels, grade and subjects.

Table (3): Overall distribution of various tests in Jordan by levels, grade and subjects

Management	MoE-DET			NCHRD			MoE-DET	NCHRD		NCHRD	Schools-MoE	MoE-USAID	
Subjects	National Test (Math, Science, Arabic, English)			NAfKE (Math, Science, Arabic)			Tawjihii (All Subjects)	TIMSS (Math, Science)		PISA (Math, Science and Reading)	School-based Assessment	Early Grade Assessment (Reading & Math)	
Grade levels	4	8	10	5	9	11	12	4	8	15 Yrs Old	All grades and in all subjects	2	3
2000	Y	.	.	.	.	.	Y	.	.	.			
2001	.	Y	.	.	.	.	Y	.	.	.			
2002	.	.	Y	.	.	.	Y	.	.	.			
2003	Y	.	.	.	.	.	Y	.	Y	.			
2004	.	Y	.	.	.	.	Y	.	.	.			
2005	.	.	Y	.	.	.	Y	.	.	.			
2006	Y	.	.	Y	Y	Y	Y	.	.	Y	Y		
2007	.	Y	.	.	.	.	Y	.	Y	.	Y		
2008	.	.	Y	Y	Y	Y	Y	.	.	.	Y		
2009	Y	.	.	.	.	.	Y	.	.	Y	Y		
2010	.	Y	.	.	.	.	Y	.	.	.	Y		
2011	.	.	Y	Y	Y	Y	Y	.	Y	.	Y		
2012	Y	.	.	.	.	.	Y	.	.	Y	Y	Y	Y
2013	.	Y	.	.	.	.	Y	.	.	.			
EXPLANATORY NOTES	NT data before 2006 was considered unreliable and misalighed by the MoE. As a result, there was no attempt to obtain the data pre-2006.			NAfKE was developed under NAfKE I program in 2003 in order to assess student performance of knowledge economy needed skills. There have been only 3 cycles.			Tawjihii is a high stake annual terminal exam that determines whether student goes to post-secondary education or not and which univeristy and, even more, which subject area to study.	Jordan participated 4 recent cycles of TIMSS, started in 1999 (the year that is not listed in this table). NCHRD is the administrator of TIMSS in Jordan		Jordan participated 3 recent cycles of PISA. NCHRD is the administrator of the PISA program in Jordan	This is considered as a continuous or portfolio student assessment. The school-based assessment result determines if a student is promoted to the next grade or repeat a year.	EGRA and EGMA were initiated by the joint agreement between the MoE and USAID to assess early grade literacy and numeracy knowledge and skills. This is only a baseline. New EGRA and EGMA assessment will take place in 2014	
Note: Y indicates that assessment took place and data is available.													

As depicted in the table above, we found that each student assessment may have its own unique cycles, scales, time intervals, and grades covered. In addition, each assessment is designed for a set of unique purposes and objectives, measuring different domains, developed by different technical experts, administered to different schools, and results analyzed and used differently. We describe each student assessment in more detail below.

IV.1.1. National Test (NT)

The NT is a census-based test developed and managed by the Department of Examinations and Tests (DET) in the MoE. The largest standardized assessment by the scale in Jordan, the NT assesses all students in Grades 4, 8, and 10 in all schools on the performance of Arabic language, mathematics, science and English language.

However, as noted in the table above, the NT only assesses a single grade in any given year and it takes three years to repeat the test grade. For example, NT assessed the 4th grade students in years 2000,

[18] EGRA and EGMA were sponsored and supported by USAID and is being considered to be part of institutional student assessment systems in the country.

2003, 2006, 2009, and 2012 (the latest). In other years, the NT skipped Grade 4 students. For the 8th and 10th grades, NT alternates in other years. The intervals are determined because of limited resources, manpower or institutional capacity, according to the DET staffers. From a design perspective, NT is supposed to provide a "grade cohort comparability" analysis. For example, grade 4 student performance in 2006 could be compared with grade 4 student performance in 2009, and again in 2012. However, in practice, test items have been changing often in the assessment tools for the three grades without considering comparability, resulting in the inability of conducting any longitudinal analysis and getting the meaningful results.

Furthermore, the current NT, given the census-based standardized nature of the student assessment, could be used to track individual students or schools. Unfortunately, the fixed interval of "every three years" to repeat a test in the same grade and frequent change of testing items without considering comparability over time with the grade cohort or tracking students and schools have made it impossible to "do the value added" of the national assessment system. As a result, NT results in a given year for a given grade has minimally been useful. The design of the NT could be significantly improved, and more on this aspect will be discussed later in the report.

Utilization of NT

To learn about the utilization of NT, we examined multiple annual reports and conducted a focus group with the NT officials and school administrators. It is evident that the report contents and template are descriptive and very similar from year to year. The reports typically include national averages, averages of field directorates, as well as averages by school types, gender and locality. Since each report only features one particular grade (Grade 4, 8 or 10), there has been no effort to present historical trend analysis and tracking the same students over time (even though it is possible to track Grade 4 students over time when they reach Grade 8). We have also observed that there has been no explanatory models used in any NT report that would help explain why some students or schools perform so poorly while others perform extremely well. Although we were explained that DET does not collect data on student characteristics or on teachers and schools, it is known that EMIS school data and teacher data (all census-based) are collected and stored in the MoE. Linking and integrating the related data for higher order analyses is a real possibility. This suggests that more and real value of the student assessment data has not been fully realized and the utilization of the NT assessment in terms of data analysis is deficient, to say the least.

We asked the representatives from the DET how NT results being used and to what purpose. To many, the use of the student assessment results means how the report is disseminated and who gets copies of the report. The DET thus informed the research team that NT annual report is given to the field directorate and schools so that they can compare to their peers and helps to inform their plans for improvement over their weaknesses and gaps. The NT results also give insight into domains or subject level knowledge (e.g., Arabic, Math, Science, etc.) although that information is more for the schools themselves than for the directorates. When asked what important decisions have been taken based on the NT performance, DET representatives honestly stated that the information has "probably informed training and planning but in practice, much has not been done." However, as one DET officer articulated

in one focus group session that "it is difficult to know the actual utilization. The NT has evolved from 2001 and DET has made many changes in NT, and so has the MoE in terms of the educational reform. The NT report is prepared annually and aggregated results are sent to all the main policy committees and directorates. Those policy bodies and field directorates must make policies or improvement plans based on all pieces of information and NT results are only a part. However, what is known is that NT results have been used in the past to figure out major and common mistakes made by students in 2006 and have helped to prepare manuals/guidelines for learning in 2009."

To further understand the NT impact, additional effort must be made to learn how other MoE departments use the NT results in their own ways and how schools and field directorates use their NT report cards for policy actions and decisions to improve learning achievement. This is not included in this study. Results from the questionnaire completed by focus group participants have revealed that most stakeholders know the NT. The least informed group was students, as 37% mentioned they did not know the assessment. However, many principals do not know how useful the NT results are or believe the results are not useful at all (4% and 22%, respectively). Those perceptions are slightly less pronounced among teachers (4% and 17%, respectively). Therefore, even if the results are disseminated by the MoE to the Directorates, qualitative results suggest that many principals and teachers still do not realize the usefulness of NT results for their practices.

As it has been expressed during focus group discussions with teachers, "The national exam is necessary for the adjustment of the curricula and the teaching strategies. However, there are no serious consequences for schools or students who fail in such tests and as a result, these tests do not receive suitable preparations, applications procedures and discussions of the results." (Focus group notes[19], Teachers, Sofana bent Hatem First School for Girls, Marka). Nevertheless, it is important to highlight that for the majority of teachers, NT results are at least somewhat useful, despite its limitations:

"*Teachers benefit from the results of the national test at the time of preparing school tests. Teachers train students on this test because it contains different types of questions which depend on multiple choice items. However, the results are not discussed during the academic year because they are sent to the school at the end of the year, and there no students at school, so the school didn't make any changes in light of the results and there wasn't interaction with the results."* (Focus group notes, Teachers, Irbid Town Prep (G/S3) for Girls, UNRWA, Irbid).

"Stakes"

NT is considered in Jordan a low-stake test for students or teachers since it does not have any "consequence" for students and teachers regardless of its performance. This is largely in line with our expectation as well as in line with MoE's expectation because NT by design is to assess individuals' performance. However, NT should have much higher "stakes" for the MoE and field directorates as NT is the only nationally required learning achievement assessment system for primary, middle and

[19] All quotes used in this report were extracted from focus group notes taken by note-takers in Arabic during focus groups and later translated into English.

secondary levels. Unfortunately, from our meetings with many stakeholders in the MoE, field directorates, schools, we have learned that there is quite low stakes or even visibility associated with the NT or NT results. Many don't know what the NT is, with many expressing "we don't know what it is and don't study for it and teachers don't teach to it." Some of the schools visited told us that they never received NT results or school report cards on NT performance. It is confirmed that no student is "punished" based on the NT performance, nor is teacher performance review linked to the NT performance, nor is the performance review of head teachers, supervisors or principals. According to most students, "we didn't care about the national exam because there weren't marks that would affect their performance in school." (Focus group notes from a student in a Secondary School). Teachers shared similar views: "…students are not ready and interested in such test as NT because it doesn't have marks. The NT could become important and relevant if it were organized and used well. Teachers should be given a chance to let student prepare for it. In the future, it might shed light on suitable teaching strategies." (Focus group notes from a teacher, Irbid). Undoubtedly, both students and teachers should be well informed about their NT performance. Should the MoE consider raising the stakes of NT by linking the NT performance to accountability measures? This is a further reasonable policy question but deserves a further pre-policy risk and benefit analysis.

IV.1.2. National Assessment for Knowledge Economy (NAfKE)

In 2003 the MoE launched a large education reform program, ERfKE, aiming to enhance the education system quality and producing graduates with knowledge economy skills. These include problem-solving skills, analytical thinking skills, computer technology skills, communication skills, etc. (NAfKE report 2007). With this vision, the MoE soon launched the new curriculum reform and development for all grades (in multiple phases) to prepare students for life-long learning and mastery of the new skills. In 2006 it implemented the first phase of the new textbooks and teaching methods (e.g. promoting student centered teaching methodology, multi-facet ways of learning, collaborative learning, etc.) and new assessment tools for Grades 1, 4, 8, and 10. The curricula change for other grades were followed in a systematic process.

As a result, NAfKE was initiated in 2006. Experts and subject specialists were brought together to develop the assessment instruments for three subjects, Arabic, Math and Science. NAfKE is a sample-based but nationally representative standardized test for 5th, 9th and 11th graders in Math, Arabic and Science subjects. Since the inception, it has been repeated in 2008 and in 2011. The purpose of NAfKE is to assess students' cognitive abilities and readiness for applying the knowledge and concepts in solving problems in real life scenarios. As part of the NAfKE system, characteristics and learning related perceptions of students, teachers and principals are also collected in each cycle, which is spaced every 2 or 3 years.[20] There has been no particular consideration to tracking the same students over time as they move up the grade ladder in the school system.

The NAfKE result in 2006 was by design to be used as an initial baseline per ERfKE program to measure the changes in learning performance before and after the new curricula, new teaching methods and

[20] NAfKE 1 was carried out in 2006. NAfKE was conducted two years later in 2008. Then NAfKE 3 did not happen until 2011, which waited for 3 years. Now, NAfKE 4 is scheduled for 2014.

learning materials under the ERfKE program. The NAfKE aims at identifying the mastery levels of skills and cognitive abilities of students from the three grades in three subject areas: mathematics, science and Arabic language. More importantly, NAfKE test comes with much more necessary information on students, teachers and schools. Not only does it support analytical effort to measure differences such as achievement by boys and girls, rural and urban schools, school authority (public, private, UNRWA[21]), and school type (discovery schools, non-discovery schools), but also it support analytical inquiries such as what characteristics and/or school factors may explain the variation in student learning achievement.

The instrument went through a thorough preparation process in selecting and piloting the test items and ended with a set of items with solid psychometric properties with high overall reliability and reasonable difficulty and discrimination levels. In addition to the test items, three supplementary tools were created to gather survey information on test takers, their teachers, and schools based on school principals. Below is the table (4) that outlines additional information collected through NAfKE over the past cycles.

Table (4), Information on NAfKE **test takers, their teachers, and schools based on school principals.**

The questionnaire	The information domains
Student	- Learning and teaching environment in the school - Attitude toward math , science and reading - Student background - Problems that facing their learning - Computer use.
Principal	- School characteristics - School environment - Principal background - ICT use - Problem facing the school
Teacher	- Teaching and learning practices. - Problems facing the teaching and learning - Teacher background - Teacher professional development - Computer use.

Change in NAfKE

The NAfKE was administered for the first time in 2006 to be one of the baseline studies for ERfKE and to be also a major indicator relied upon in evaluating ERfKEI and ERfKE II during the years (2003-2009) and (2010-2015).

[21] UN Relief and Works Agency schools for Palestine refugees.

In 2008 NAfKE was administered for the second time with the same cognitive items and questionnaires items, whereas in 2011 minor changes on the questionnaire items were made; for example, the formatting was slightly changed and some items were added to enlarge coverage.

In 2013 a joint team from the MoE and NCHRD worked in reviewing the cognitive items as well as the questionnaires items on the NAfKE, and the technical team reconsidered the cognitive and the knowledge economy skill weight. Consequently an item pools for math, science, and Arabic were developed and piloted, in addition to that the questionnaires has been reviewed to increase coverage and to get more reliable results. NCHRD will administer the revised NAfKE in 2014.

Analysis used in NAfKE

In all NAfKE cycles (2006, 2008, and 2011), the descriptive statistical methods such as means, standard deviation and percentages were used. In addition, some inferential statistical methods—such as T-test for independent samples, Analysis of Variance (ANOVA), and regression analyses[22] were used.

Coding and data entry for NAfKE

As we indicated earlier, NAfKE cognitive items included open-ended questions as well as multiple-choice questions. NCHRD develop a coding guide for mathematics, science, and Arabic. Supervisors specialized in these subject domains were trained to use these guides, so they code the items accordingly.

MCQ entered to the computer directly and items keys are used to score the items. The software WinDEM is used as entry software and also is used to clean the data. However, the data entered to the software through a selective list of data entry persons whom are very well trained on the software and they are quick with approximately no errors.

Although NAfKE has existed for several years and it has been applied in many schools, overall knowledge about the assessment was low among all focus group participants. Results from the questionnaire suggest that 88% of teachers, 86% of students and parents, and 46% of school administrators did not know what NAfKE is. Those results were highlighted during focus group discussions:

"…Parents are aware their children participated in TIMSS and PISA but they don't know NAfKE".(Focus Group Notes, Parents, Al- Rusaifa the Third, School for Boys, UNRWA).

"The teachers know that the national exam is important and it gives a feedback for the students and the teachers. They were also aware of and had experience with TIMSS and PISA, but they didn't know about NAfKE." (Focus Group Notes, Teachers, Al- Rusaifa the Third, School for Boys, UNRWA).

"NAfKE is not known to students, so they didn't talk about it." (Focus Group Notes, Students, Irbid Town Prep (G/S3) School for Girls, MoE).

[22] We must note that regression models were developed only within each cycle of the NAfKE test. The researchers were not able to develop regression analyses with multiple years of data at individual student level due to an inability to track randomly sampled students across cycles over time.

As a result of the general lack of knowledge about the assessment, only a small number of teachers (n=13 or 14%) could report on how useful NAfKE results were for themselves and for students. Among those who knew NAfKE, most reported that the assessment was either somewhat (n=5) or very useful (n=5) for themselves, as teachers. A similar number of teachers reported that NAfKE results are somewhat (n=6) or very useful for students (n=3) as well. Among the school administrators who knew NAfKE (n=15 or 54%), two-thirds thought the results were somewhat or very useful for administrators and teachers. A smaller number (n= 8) thought the results were useful for students. This evidence suggests that NAfKE be taken more seriously in order to reach the intended effectiveness. On one hand, it is understandable that NAfKE should be kept as a relatively low stake test for students and teachers at school level as it is likely to reflect a real "truism" of the learning performance. On the other hand, the NAfKE test must be elevated to a higher stake of policy importance by conducting policy-relevant analysis at national and regional levels and sharing the analytical results widely.

IV.1.3. Tawjihii

Article (29) issued in accordance to the Education law No. (3) For the year 1994 stated the following: "The MoE conducts a general exam for the students by the end of the secondary level in overall curricula of secondary education. A successful candidate is awarded a certificate called General Secondary Certificate." According to a working paper prepared by the MoE entitled "Developing General Secondary Exam," the purposes of the Tawjihii exam are to demonstrate the level of learning that has been gained by students after the completion of the secondary level, the effectiveness of teaching and learning process, and abilities to continue the tertiary education.

However, the *Jordan Times* (April 8, 2013) published a commentary: "Unlike the globally recognized "SAT" and the A-level exams, Tawjihi acts as an irrefutable verdict in university admission. Tawjihi has a huge impact on the students, and in order to perform well, much attention is given to Tawjihii, at the expense of time allotted to critical thinking. High school teachers give their students the best opportunity to succeed in college admissions by teaching them rote answers for exams, which does not challenge their critical thinking. Teachers have to work outside the system to promote their students' development by giving private lessons; students are just well prepared to answer questions in rote formulas."

Change on Tawjihii

In recent years the MoE implemented many procedures to further develop and improve Tawjihi, including:

- MoE conducted a study to identify the student views about Tawjihi;
- MoE collected observations and comments from teachers, principals and students, as well as from the concerned parties inside and outside the Ministry;
- MoE developed a table of specification for each subject through teachers' and supervisors' participation;
- MoE prepared working papers by experts and professors to develop the exams. Consequently, the MoE developed multi scenarios to enhance Tawjihii.

In 2013/2014, the MoE decided to remove the multiple choice items from the exams papers[23]; all questions became essay questions. In general, the development of general secondary exam has been focused only on developing the cognitive domains measured by the Tawjihi exam and the weights of these domains.

Given the high stakes associated with the Tawjihii exam for students, it is not surprising that focus group participants believed that the assessment was important. For example, most students rated the Tawjihii as either very important (89%) or somewhat important (6%). Those perceptions were shared by teachers, parents, administrators, and university faculty members. However, many other focus group participants (non-student participants) had conflicting views about the Tawjihii. For example, among university professors, five (out of seven) believed that Tawjihii results were not very important in predicting students' academic success. Only two faculty members believed the results were somewhat important:

> "*Tawjihi in 1964 was used as the system to evaluate students and send them to respective majors. It is not a working system, it has failed…there are violations on Tawjihii from students, parents, and local community…It is fairly easy for experienced teachers to guess the questions that will be on the exams. Our tests don't encourage higher level thinking. Tawjihii is an achievement test, it does not measure reasoning and critical thinking."* (Focus Group Notes, University Professors, Amman, February 4, 2014).

Parents, teachers, students, and administrators recognize, for the most part, that Tawjihii is important and that results might determine a child's academic and professional future. However, most stakeholders expressed concerns over the limitations of Tawjihii and the burden associated with the test for all education stakeholders:

> "*Tawjihii is important and necessary, but it may cause problems and psychological pressure because the result of this test determines the future of the student. There are many subjects covered by Tawjihii and these subjects require a lot of memorization."* (Focus group notes, Parents, Irbid Town Prep (G/S3) for Girls, UNRWA, Irbid)

> "*Tawjihii is an important exam to join the university, but it is like a terror movie that all people watch (and live in). People have lost trust in it."* (Focus group notes, parents, Queen Zein Elsharaf School for Girls, Aqaba).

In summary, Tawjihii must reform itself. Based on how people in Jordan think of the Tawjihii today in an overwhelmingly negative way, it deserves a serious policy attention at the highest level of the government. MoE must take an unprecedented measure to reform the Tawjihii and make it work for the best interests of the educators and the educated.

[23] It is not known to the researchers why multiple choices were moved.

IV.1.4. TIMSS

TIMSS is an international student assessment managed by the International Association for the Evaluation of Educational Achievement (IEA) to test to assess student achievement of Grades 4 and 8 throughout the world and to allow for country-to-country comparisons. As stated by the IEA, TIMSS helps participating countries "share the conviction that comparing education systems in terms of their organization, curricula, and instructional practices in relation to their corresponding student achievement provides information crucial for effective education policy-making."[24] In general, participating countries also use TIMSS in a variety of ways to explore educational issues, including: monitoring system-level achievement trends in a global context, establishing achievement goals and standards for educational improvement, stimulating curriculum reform, improving teaching and learning through research and analysis of the data, conducting related studies (e.g. monitoring equity or assessing students in additional grades), and training researchers and teachers in assessment and evaluation.[25] In 2011, the last TIMSS cycle, 45 countries participated in the assessment of 8th graders, and Jordan was ranked 28th in Science and 35th in Math. Table (5) presents TIMSS test included subjects and main domains.

Table (5): Information on TIMSS test included subjects and domains.

TIMSS Test Subjects and Domains:		
	Content Domains	**Cognitive Domains**
Math	Numbers	Knowing
	Geometry	Applying
	Algebra	Reasoning
	Data & chance	
Science	Life science	Knowing
	Physics	Applying
	Chemistry	Reasoning
	Earth science	

Jordan's participation in TIMSS put Jordanian student learning performance on the world map. It provides important comparative information at a global scale on Jordanian students' learning and academic achievement.

IV.1.5. PISA

Launched in 2000, PISA is also an international student assessment system. PISA is managed by the Organization for Economic Cooperation and Development (OECD) and administered every three years to 15 year-old students across countries. The primary objective of PISA is "to determine the extent to which young people have acquired the wider knowledge in reading literacy, mathematical literacy and scientific literacy that they will need in adult life."[26] In each round, one specific domain is taken as the

[24] (http://timss.bc.edu/timss2011/index.html)
[25] Ibid.
[26] (OECD, 2004)

main subject, occupying about two-thirds of the testing time, with the remaining testing time being divided between the other two "minor" domains. Thus in 2000, the main focus was reading literacy, in 2003 mathematical literacy, and 2006 scientific literacy. OECD also claims that PISA "provides insights into the factors that influence the development of the skills at home and at school and examines how these factors interact and what the implications are for policy development."[27] PISA test subjects and domains are listed below as presented on Table (6).

As regards stakeholders' level of awareness about TIMSS and PISA, questionnaire findings suggest that most stakeholders do not know those two assessments, even though Jordan results and ranking in relation to other countries are highly publicized by the media. School administrators seems to be the most knowledgeable, with only 21% and 18% reporting that they do not know either TIMSS or PISA, respectively. Students seem to be the least aware about those assessments, with 83% reporting not knowing PISA and 70% not knowing TIMSS. Teachers' level of awareness about TIMSS and PISA were also low at 55% and 40%, respectively.

Table (6): Information on PISA test included subjects and domains.

PISA Test Subjects and Domains:		
	Content Domains	**Cognitive Domains**
Math	Space and Shape	Formulating
	Change and Relationship	Employing
	Quantity	Interpreting
	Uncertainty	
Science	Physical systems	Identifying scientific issues
	Living systems	Explaining scientific phenomenon
	Earth & space systems	Using scientific evidence
	Technology systems	
	Scientific inquiries	
	Scientific explanations	
Reading	Continuous text	Access and retrieve
	Non-continuous text	Integrate and interpret
		Reflect and evaluate

"*International tests (TIMSS and PISA) are not known [by teachers], so they cannot talk about their importance, benefit and necessity*." (Focus group notes, teachers, Irbid Modern School, Co-Ed, Irbid, [insert date]).

"*My colleagues don't have enough knowledge or are fully aware of these tests [TIMSS, PISA, NAfKE]. We are drowning with teaching education, especially Tawjihi and school-based tests.*" (Focus group notes, faculty members at University of Jordan, Amman, February 4, 2014).

[27] (OECD, 2003)

Among stakeholders who knew one or both assessments, there was awareness about their limitations with regards to awareness of test results, their usefulness to inform classroom practices, low "stakes" associated with them, and procedures at the school and/or directorate levels. Those concerns were apparent during focus group discussions:

> "*Tests such as TIMSS and PISA include items that focus on analysis [utilization of knowledge]. Most of the teachers do not expose their students to experience based on analysis- moreover, schools do not reach beyond the level of [basic] knowledge. "(*Focus Group Notes, Supervisors, North).

> "*Teachers did not benefit from NAfKE, TIMSS and PISA test results tests because the results were not sent to school. So there were not discussions, interactions, or taken procedures in the light of the results.*" (Focus group notes, Teachers, Irbid Town Prep (G/S3) for Girls, UNRWA, Irbid)

> "*There are no consequences if students don't do well in those exams (TIMSS, PISA, NAfKE). Students do the exam and that's it. Neither the students nor the schools are informed about students' results. A lot of students are not trained for such exams, because it is not from their curricula and the results are not important to them."* (Focus group notes, Teacher, Faisal Al-Awwal School for Boys, Aqaba)

IV.1.6. School Assessments

School assessment in Jordan is a portfolio and year-long on-going assessment that considers the results of local teacher developed tests and quizzes including mid-term and end-term tests and student attitude and behavior in the classroom. Although this assessment is well informed by the MOE's assessment framework, implementation guideline and rubric, schools and teachers developed their own interpretations and practices for the assessment purpose. It is generally expected that there will be overall inconsistency across all schools and field directorates in terms of levels of difficulty in weekly quizzes or term examinations or how students are expected to behave in the classroom. However, each student gets an overall score composited from the various components and subjects at the end of school year. Individual teachers have the authority over the given final score in a given subject. Upon the final score, a pass or failure is determined based on the MoE's rubric and guideline to either promote student to the next level or repeating a grade[28]. According to the DET, the school assessment framework is used by both public and private schools. The national standard rubric for the school assessment is developed by the DET.

The Jordanian school year is separated into two semesters, each lasting four months. During the semester, a student is assessed in the following ways: three assessment components, each comprising 20% of his or her grade, and a final examination comprising the remaining 40%. The three components, according to the DET personnel are distinct from each other: 1) the materials learned over the first two months of the semester, and student achievement is assessed through paper-and-pencil midterm

[28] A policy exists in the MOE that students from the first 3 grades do not repeat. Once a student repeats a grade, he or she will not repeat the same grade for the 2nd time. Over all years of schools no students will repeat more than twice.

examination; 2) short exams and quizzes throughout each semester; and 3) student's overall performance over student portfolios, checklists, and classroom observation, amongst other tools (by subject teachers). The table(7) below sums up the grading rubrics for teachers.

Table (7): Classroom Assessment Guidelines from MoE for Teachers

1st Component/Period (20%)	2nd Component/Period (20%)	3rd Component/Period (20%)	Final End of Term Test (40%)
Paper and Pencil Mid-term Exam.	Group of Short Exams and Quizzes	Student Portfolios, Checklists, Classroom Behavior, etc.	Final Exam

Based on the MOE's framework for school/classroom assessment, the implementation focus should be on five specific strategies:

1) Paper and pencil assessment: written quizzes and tests.
2) Performance based assessment.
3) Self-evaluation assessment: student portfolio.
4) Classroom observation: checklists, rubrics, rating scales.
5) Interactions with other students and with teachers.

These general strategy areas are all part of the MOE's framework template but there is flexibility built into the system, according to the DET officials, to allow teachers to adapt these strategies to their classrooms and to enable them to necessarily manage how they perform the assessments. Teachers can choose to employ different tools and different strategies to determine a student's final mark. Furthermore, the tools advocated by the MoE can be employed for multiple strategies and are not necessarily specific to only one.

The records for classroom assessment are kept in both hard copy and electronically. They are kept with the schools until the end of the year and then they are entered into a computer database, EduWave, which is managed by the Queen Rania Center. Once the data is entered, the MoE would have the data accordingly. It is reported that the Queen Rania Center has all the school assessment data for the past few years since the inception in 2004[29]. Paper copies are only kept in the schools and a duplicate is submitted to the MOE.

Utilization of School/Classroom Assessments

School/classroom assessment is designed to determine individual student performance in school/classroom and decide if he or she will be promoted to the next grade or repeat the same grade.

[29] As part of the ERfKE I program, a new school assessment (school-level continuous assessment) was launched in 2004. It was implemented in several phases. By 2007 all schools are mandated to apply the framework and rubrics. Over the years, rubrics for the final assessment score have been changed slightly in terms of weightings for each component. But the major components remain the same.

Under the assessment scheme, a 50% point in the final mark is the pass/fail "cut score." If a student scores below 50%, he or she may have to repeat, depending on a subject from the 4th grade on. For example, if a student fails Math, the student must take a supplemental exam/incomplete test for Math. If he or she fails the supplemental exam, then the student must repeat the entire grade. If a student fails more than three subjects in the first place, the student does not have an opportunity to re-take any exam and will repeat the grade.[30]

There is usually no policy action taken based on the school assessment results at the field-directorate level or the central MoE. Although data is annually submitted to the MoE (the Queen Rania Center) and shared with the field directorate office, there has been no aggregated analysis or reporting on school assessment. It is possible that the MoE and field directorate could conduct necessary school level analysis by connecting school/classroom assessment results with the NT results. This would not only allow the MoE to learn how the two assessments can be mutually validated between each other but also identify the schools that may need significant support and resources most urgently in order to enhance the performance level. If both standardized NT and non-standardized classroom assessment results confirm the worst performance levels are in the same group of schools in Jordan, the MoE would be more confident and have better ideas in terms of which schools must be targeted.

"Stakes"

In general, school/classroom assessment has the highest "stakes" for students and parents in Jordan. This is confirmed by interview results with students, teachers and schools. Most students when asked about the importance of all types of assessments, they often point out that it is school "report card" they pay attention to or worry about. The report card is given to student at the end of the year to inform the results of school assessment after one academic year and decision on promotion or repetition. The stakes are high for students since the performance result has a considerable consequence after one year. Repetition is considered a shame and an ultimate punishment for students (and even parents) according to the DET officials. Many students and parents work hard to ensure their "passing" scores. Given the "flexibility" teachers and principals may have in the school assessment, stories of "negotiating" for passing grade between parents and teachers are often heard throughout the system, according to one MOE staff member.

According to focus group discussions, 88% of students who answered the questionnaire asserted that school-based assessments were either somewhat or very important for them, especially in comparison to other international and national assessments. According to some students, "students take school tests seriously. The most important tests are the written ones [as opposed to oral examinations] because students have a chance to write freely without fear. Students did not take NAfKE and TIMSS tests seriously. (Focus group notes, students, Irbid Town Prep (G/S3) for Girls, Irbid).

[30] Student may repeat only from 4th grade up. However, no student is permitted to repeat twice in the same grade according to a MoE policy. Throughout student career, any student can only repeat twice regardless how many times he or she fails. In addition, schools are capped by the number of student failing grades. According to the MoE, no school should fail more than 10% of students in any grade.

"Students look at school assessments more seriously and think they are more important than the national and international exams. (Focus group notes, students, Al- Kindy Secondary School for Boys, Marka).

The importance of those assessments for teachers and parents was also apparent. 90% of teachers and 75% of parents rated school-based assessments as either somewhat or very important. Those results are understandable, as teachers rely on results to identify students with difficulties, shape their instruction accordingly, and ultimately decide who will be retained in the same grade for an additional year.

IV.1.7. EGRA and EGMA

EGRA and EGMA assessments are developed to assess students´ learning in early primary grades and their mastery of foundation skills upon which all other literacy and mathematical skills can be built. EGRA and EGMA also intend to ascertain key school characteristics and components that can foster learning. An additional tool, the School Management Effectiveness (SME), provides a comprehensive assessment of school and classroom characteristics traditionally associated with pupil performance. EGRA and EGMA are initiatives supported by USAID/Jordan in partnership with the MOE, and implemented by RTI International under the Education Data for Decision Making (EdData II) project to conduct the SSME. EGRA and EGMA are sample-based assessments and are supposed to provide a baseline of the current situation regarding literacy and math skills in Grades 2 and 3. An SSME survey is used to interview school principals and teachers, to conduct inventories of school and classroom resources, to observe reading and math lessons, to inform future education policy decisions. The instruments used during the baseline —the National Early Grade Literacy and Numeracy Survey in Jordan—were adapted specifically for the Jordanian context in cooperation with staff from the MOE[31].

Although EGRA and EGMA has not been part of the overall MoE assessment yet, it is important for MoE to adopt it institutionally and use it as "early-stage detection" mechanism to identify specific learning needs, ineffective teaching and other related impediments to quality education and to support focused and remedial improvement programs timely.

IV.2. Utilization of Student Assessment Results in Jordan

Earlier, we discussed the issue of utilization in each of the student assessments in Jordan as we described each assessment system. It is the issue of such great and critical importance that we decided to further discusses the issue in its own sub-section. We want to focus on the overall utilization of all assessment data and results. For this, we examine two different types of utilization of student assessments, 1) utilization of the raw data of student assessments for technical analysis and 2) utilization of the assessment results with regard to students, teachers, schools, field directorates and national education system for policy or decision actions.

[31] Aarnout Brombacher, Penelope Collins, Christopher Cummiskey, Emily Kochetkova, and Amy Mulcahy-Dunn (2012). Student Performance in Reading and Mathematics, Pedagogic Practice, and School Management in Jordan. RTI

The Mapping Study examined these issues and has made a general conclusion that the actual utilization of the raw data for technical analyses (research or evaluation analysis) in Jordan is insufficient and the utilization of the assessment results for policy decisions or improving quality is negligible. Purposes and objectives of student assessments are clearly stated in the student assessment frameworks in Jordan. Each has unique and important purposes and objectives. But these broad objectives do not inform how raw data should or could be analyzed, nor how assessment results should or could be used.

IV.2.1. Utilization of Raw Data

While it is understandable that there is no guide book for using the raw data[32], it must be recognized that proper use of the raw data, often termed as data analysis, often leads to better chance for using the results by policy makers (World Bank 2009). Having examined multiple years of reports on NT, NAfKE, TIMSS and PISA performances and interviews with testing administrators, field directorates and school stakeholders (principles, teachers and students), we realize that the use of raw data is limited. First, raw data is rarely shared or analyzed by others, even within the education sector in Jordan. Secondly, data from multiple years, multiple levels of the education administration and multiple sources within the education sector are not systemically integrated at the original raw data level, which result in an incapability to carry out higher order data analyses to identify key education problems and relevant factors that contribute to the problems, or clarify policy implications based on the higher order data analyses.

Data sharing is inadequate

We discussed data issues extensively with DET officers and NCHRD researchers. It became clear to us that raw data was by default not shared openly (e.g. openly available on the MoE or NCHRD web site) unless there is "an official request" as one MoE officer explains. This is particularly true in the MoE. NCHRD shared some of its TIMSS and PISA assessment data with local universities in CD packets given the fact that international TIMSS and PISA data by countries is publicly available (downloadable from TIMSS or PISA websites). But in general, there has been limited data sharing. Many countries in North America, the European Union and developed nations in Asia start to make raw data available online and downloadable for secondary data analysis and on-going use.[33] With the limited data sharing, one can only conclude that there has been no or limited secondary data analysis of student assessment results in Jordan. The MoE uses or analyzes NT data and NCHRD uses and analyzes NAfKE, TIMSS and PISA data. Perhaps the demand from other stakeholders is not present or perhaps there is a hesitation to make the raw data available by default for any further use. In the Middle East, Jordan could start to champion the initiative of developing an open and transparent data system in education sector.

For the Mapping study, researchers requested NT data from the MoE and NAfKE data from NCHRD, and the data was provided even though the requests were considered as the "first requests" relatively.[34].

[32] Raw data here means that original data, coded and cleaned, in its original form of unit records.
[33] We must note that in almost all cases in the world, the official downloadable data would conceal individual (student, teacher or school) identity. Raw data is only used for secondary data or statistical analysis.
[34] We must note that sharing individual NT performance data does not have to violate individual student confidentiality. For example, for this study, we observed individual NT data but no individual name or personal id is requested.

This confirms that there could be "demand" problem[35]. The demand problem may come from lack of assessment related inquiries, lack of higher-order research questions, and/or lack of technical analysis skills. Regardless, we strongly believe that the demand be stipulated and increased and data inquiry capacity enhanced and advanced so that data sharing culture cultivated and nurtured.

Data analysis is insufficient

Over the past years, the MoE produced annual NT performance report (national and directorate level). NCHRD produced NAfKE, TIMSS and PISA reports in each of the cycles. We will discuss Tawjihii and school assessment separately. There has not been much difference from year to year in the contents, methods, or format of the reports. For NAfKE, TIMSS and PISA reports, we could see that technical or statistical skills were used since they are all sample-based tests such as mean comparisons (T-testing and Analysis of Variance, etc.), composite development (reliability testing, principal component analysis, etc.), and explanatory models of multivariate analyses (regressions, etc.). However, these reports lack policy relevance and focus. They tend to be too long and often overtaken by statistical tables with minimum "policy stories" to report. Statistical techniques were used but not always in the best or the most needed way in terms of addressing policy issues or problems. Presentations of the statistical analysis results with narrative explanations are not quite reader-friendly[36]. Readership of the reports is very low according to the recent study. Recently, policy briefs were written by NCHRD based on the latest NAfKE, TIMSS and PISA reports in order to reach larger audience and increase the visibility of the acute problems identified from the reports.

For NT report, 80% of the report each year simply covers descriptive tables or bar charts with little narrative, only reporting on averages or percentages of NT performance levels by school type, gender, and field directorate.[37] Although the NT annual report only targeted domestic educators (more likely for internal users with no English version), the annual report does not serve well to monitor student learning performance at the system level; does not identify real needy schools and students who may need more support or investment; and does not evaluate how satisfactorily students performed at national, regional and school levels. The simple averages presented fail to tell the "stories" of the NT performance levels, and the annual report therefore fails to capture the usefulness of the assessment data for informing policies.

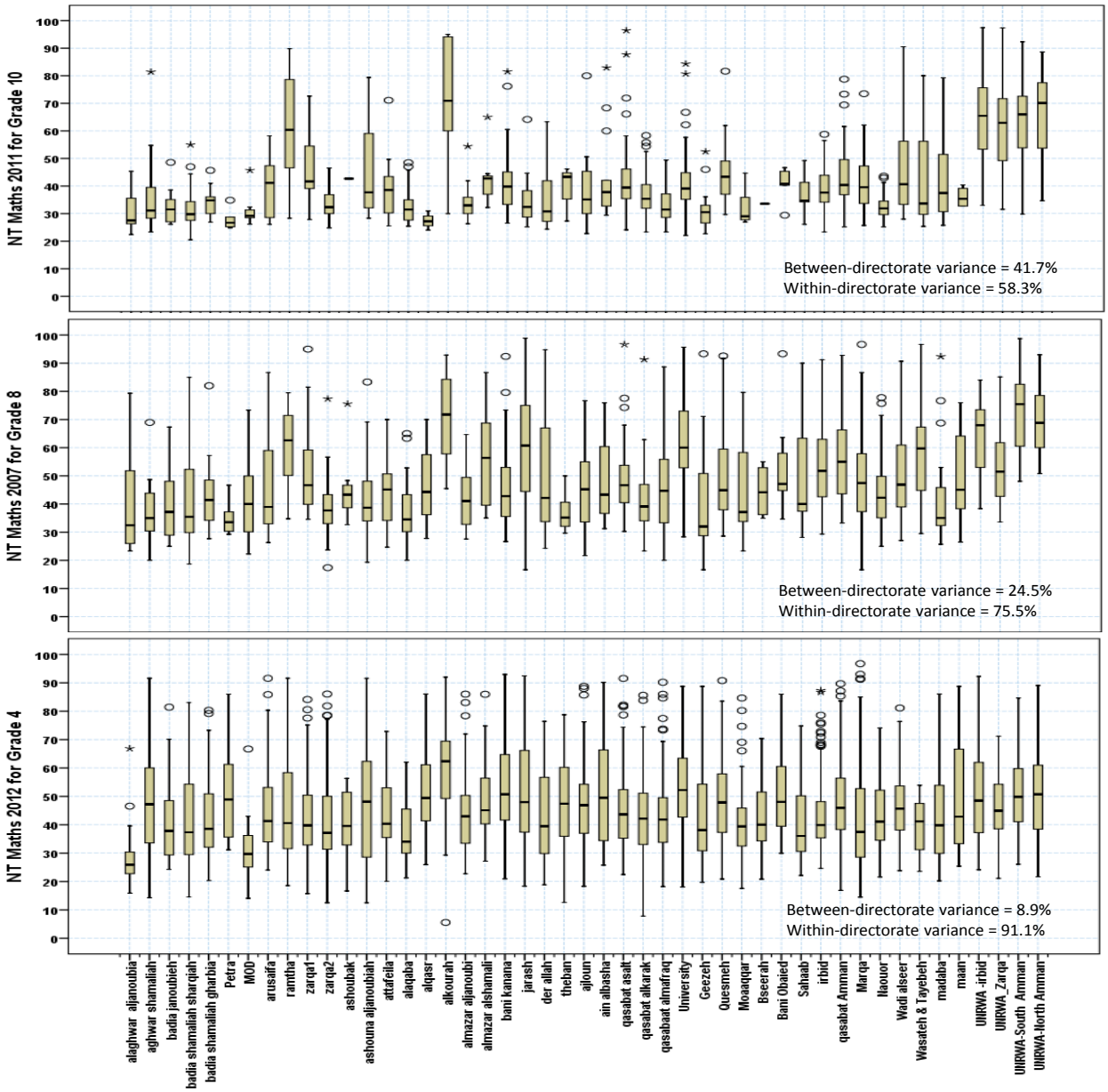
As an example, the chart below presents an interesting Illustration-proven possibility to identify a hidden problem of Jordan education system from NT data analysis we conducted. Since NT is administered only one single grade each year, we managed to examine the performance levels for all three grades from three years. For example, we obtained NT Math data from the MoE on Grades 4, 8, and 10 in the years of 2007, 2011, and 2012 and constructed "system performance maps" model to

[35] Directorate of Examination and Testing (DET) provided student NT raw data for the years of 2007 (grade 8); 2011 (grade 10), 2012 (grade 4)

[36] Although original reports are in Arabic language, translated English versions provided evidence for us to draw the conclusion.

[37] We fully understand that NT is census-based student assessment. When data is analyzed, there is no need to follow stringent inferential statistical norms or rules. However, certain type of descriptive statistical analysis from systems perspective is critical.

analyze the between directorate and within directorate variances to see if there is any hidden problem (see below). Three "boxplots" (stacked together) represent three different grade level NT Math performances (Grade 4 shown at the bottom, Grade 8 in the middle and Grade 10 at the top) by field directorates. As one may notice that within field directorate, the best performing students (25%) are indicated as a line above the box and the worst performing students (25%) are indicated as a line below the box. The middle level performing students (50%) are indicated as inside the box.



The three boxplots are "national maps" locating student Math performances for the representative grades in primary, middle and secondary levels in Jordan. They look quite different from each other. The first boxplot from the bottom (NT math performance of Grade 4 students) shows that there are large variances within most of the field directorates (shown as long lines and boxes). But the differences

between directorates are relatively small. In other words, the directorates are somewhat identical or homogenous. Each directorate has a group of excellent performing schools and students as well as poor performing schools and students. This is clearly indicated in the statistics called between-directorate variance component (8.9%) and within-directorate variance component (91.1%).

However, as for the Grade 8 student NT math performance, we see that the between-directorate variance goes up dramatically to 24.5%, which indicates that there is significantly increased variation component (almost 3 folds of the Grade 4 level) between field directorates in Grade 8 student NT math performance. Furthermore, when we examined the Grade 10 student NT math performance, we find that significantly increased again is the between-directorate variance component, now up to 41.7%. With these statistics, Jordan could be ranked at a top in the world having one of the highest between-district variance components within a country (A. Riddell 2002). Why students' learning performance varies dramatically among directorates as students move up to higher grades in the Jordanian education system remains a puzzling question. This question deserves a serious policy attention and in-depth research and evaluation analysis, which is beyond the scope of this study.

Although this is only an illustration to demonstrate how NT data could have been used to identify the system's hidden problem(s), we hypothesize that there is a large variation in local capacities (field directorates and schools), of school management, quality of teachers and school environments which could have strongly correlated with the student learning performance. As students move up to higher levels, the accumulated negative effects of the poor local capacity in management and education quality get exacerbated, causing an "unforgivable damage" in student learning achievement in later grades in Jordan. This is surely preventable, particularly if the systemic problem is detected early through the right analysis of the existing data. To again emphasize the point, we strongly believe that there should be more higher-order analyses carried out with the NT data. The trend seen here in dramatically increasing the between-directorate variance component from Grade 4 to Grade 10 in NT Math performance is also witnessed in other NT assessed subjects including science, Arabic and English.

In addition, we integrated the student assessment data from NT with NAfKE, TIMSS, and PISA and in multiple years by matching the unique school code. As we mentioned earlier, we could only match any sample-based selected schools for NAfKE, TIMSS or PISA to NT schools because NT schools are all "population." However, we must note that data from participating schools in NAfKE, TIMSS, and PISA cannot be integrated because they are randomly selected and they are different schools in the country. In other words, no schools or very few schools participated in more than 2 sample-based student assessments. With the limited integration, we were able to generate correlations between NT results and NAfKE, TIMSS or PISA (paired correlations). Table (8) presents these results as on next page:

Table(8): Correlations between NT results and NAfKE, TIMSS or PISA (paired correlations).

TIMSS Scores (2007) ***(101 schools matched)***	***NT Scores (2007)***		
	Math (grade 8)	Science (grade 8)	Arabic (grade 8)
Math (grade 8)	**0.54 (p=0.000)**		
Science (grade 8)		**0.58 (p=0.000)**	
PISA Scores (2012) ***(86 schools matched)***	***NT Scores (2011)***		
	Math (grade 10)	Science (grade 10)	Arabic (grade 10)
Math (15 year old)	**0.34 (p=0.001)**		
Science (15 year old)		**0.44 (p=0.000)**	
Arabic (15 year old)			**0.48 (p=0.000)**
NAfKE Scores (2011) ***(37 schools matched)***	***NT Scores (2011)***		
	Math (grade 10)	Science (grade 10)	Arabic (grade 10)
Math (grade 11)	**0.39 (p=0.019)**		
Science (grade 11)		**0.22 (p=0.192)**	
Arabic (grade 11)			**0.45 (p=0.000)**

From the table above, it is evident that there is a significant and consistent correlation between NT and TIMSS, PISA or NAfKE results in each of the subjects and in grade 8 or 10 (except for that between NT science in grade 10 and NAfKE science in grade 11.) This generally informs us that if schools don't perform well in NT, they won't be likely to perform well in TIMSS, PISA or NAfKE. Given the low stakes of these tests for students, the results may be reflective of the reality of the state of the education sector in Jordan.

Technical Analysis Skills

According to the DET staff in the MoE, there has been limited analytics capacity in the directorate and the MOE—even though technical trainings and workshops were organized in order to analyze student assessment data through Item Response Theory (IRT), multivariate analysis, item difficulty indexing and banking, reliability tests and so on. These are indeed necessary skills, but these trainings for individuals have not been translated into the institutional capacity of the DET. Without increasing the DET capacity to produce the right kind of assessment reports, the DET will be unable to adequately evaluate the education system performance. Nor will other directorates—such as policy planning and M&E directorates—be able to integrate and analyze student assessment data with other data sources to address policy issues. Technical skills can only be useful if they can be incorporated into job responsibilities and task performance. Training can only be useful if there is a follow up coaching and technical assistance so new skills can be imbedded into real job functions and production.

NCHRD has stronger technical analysis skills noticeable in their student assessment reports (NAfKE, TIMSS and PISA) in terms of types of statistical analysis they conducted. However, there is a lack of

capacity in addressing education problems from the sample-based assessment results and from the system analysis perspective.

IV.2.2. Utilization of Student Assessment Results

Worldwide literature supports that student assessment results may be used for four purposes: 1) identify student learning needs and find ways to improve; 2) certification, progression and/or sorting or streaming students to different tracks; 3) system monitoring and evaluation; and 4) accountability. We believe that in Jordan student assessment results are generally considered for the first three purposes to some extent, but not the fourth. For example, NT is considered to monitor the national and sub-national curricular learning performance of Grades 4, 8 and 10; school assessment is known for diagnostic or progression purposes; and Tawjihii is known for certification of graduation and qualification for college.

IV.2.2.1. Identifying student learning needs and take actions to address the needs

Two most important objectives of student assessment are 1) to identify individual learning needs and provide teaching needs accordingly, and 2) to understand the quality of an education system as a whole in the historical, cross-unit and criteria-based comparative context and plan policy actions to improve the quality. In Jordan, school/classroom assessment system is clearly designed for the former and NT is for the latter. There is no question that teachers are well informed of each student performance level through the school/classroom assessment tools. Quizzes, term-exams, and classroom behavior checklist would provide undisputable evidence throughout each academic year. But how an average teacher should use the assessment evidence to address the learning and teaching needs remain unanswered. For example, should a teacher use the assessment results to self-evaluate her or his weekly teaching performance in terms of student learning and inform the next week lesson plans accordingly? Should a teacher provide extra help or supplementary classes to those who perform poorly in her or his class? Although these inquiries are beyond the scope of this current report, they are surely relevant inquiries to see if the key objective of the school/classroom assessment is soundly met. Just reporting on how students perform by methodically sorting them out with final assessment scores and decisions on promotion or repetition is measurably inadequate. We strongly suggest that there be an investigative study on the impact of school/classroom assessment on teaching and learning. We suspect that taking productive actions based on the assessment results to improve teaching and learning—and providing additional support and help—is rare and needs to be centrally supported. For more, one may examine a leading research-based approach to improving teaching and learning informed by classroom assessment results, named “Datawise Approach”, originated at Harvard University.[38]

IV.2.2.2. Graduation and Admission

The quality “graduates” from the school system may be assessed by the final assessment for a qualification purpose. Criteria for admission to the next level of education may also be added. This is often the case in large education systems such as China, Vietnam and Russia. The Tawjihii test in Jordan is clearly meant to serve the dual purposes. Students in Jordan, particularly after the 10th grade, are constantly reminded that they must do well in the final Tawjihii test or their student career will end without any further tertiary education. According to all stakeholders, Tawjihii is “brutal,” “ruthless,” and

[38] Link: http://isites.harvard.edu/icb/icb.do?keyword=datawise

"all things about the student's future." Although students may have a few chances to retake the test if they fail it initially, it is absolutely the most important test if students intend to go to a college or university. There is no other use according to the test officials. The utilization of Tawjihii results are to rank in order all graduates in Jordan in the given year and to determine an official cut-score for admitting students to universities in Jordan.

Most education stakeholders have "negative" perceptions against the Tawjihii in terms of the way that it is utilized. But there has been no alternative established to replace or supplement it. Even when we discussed the Tawjihii and the university admission policy with university professors and officials, many of them expressed their dissatisfaction with the Tawjihii's sole purpose. Quite a few expressed that "the Tawjihii system is too old fashioned" to meet the 21st century society and knowledge economy. Why the issue has not progressed past debate for so long is not known. Clearly the real debate is on the use of the results and for what purpose, not whether or not Jordan should have the Tawjihii test at all. For this reason, we suggest that "stakes" should be lowered by the authority.

IV.2.2.3. Monitoring Education System Performance

This is the most critical purpose of any national standardized student assessment, particularly when a national standardized curriculum is established in a country. Policymakers managing a national education system must know well the system-level output performance—which they can use to continuously improve the quality of the system. The NT system managed by the MOE appears to serve this purpose and so does NAfKE, as managed by NCHRD. In terms of monitoring system performance, one must apply the proper methods properly to address potential system problems. System problems may not necessarily be revealed if systems analysis is not applied (Kershaw and McKean, 1959; and Senge, 2000).

Although the purposes of the NT and NAfKE are relatively explicit in their frameworks and designs, the practical analytics demonstrated in their related reports are insufficient. It is difficult to assess if the lack of utilization of the NT or NAfKE results is a result of a lack of good systems analysis, but we believe there is a relationship between them. Policy makers in the MoE, for example, must "cope with" system problems that must be revealed with systems analysis methods based on system-level data or indicators. If there is a lack of the system data evidence or lack of systems analysis methods, there would be no revealing of systems problems and therefore there would be lack of utilization of the report.

In other words, the objectives of the assessments are insufficiently met. Although routine reports are produced and various aggregates of average scores or percentages of competent performers are rank ordered, the reports clearly demonstrate insufficient analysis and policy relevant implications. The following table(9) summarizes the level of analyses, reporting, and dissemination of the student assessment data. Five letters are used to indicate various levels. For example, F (Full) indicates that there has been extensive data analysis conducted, assessment results reported or disseminated respectively. P (Partial) indicates that data analysis partially carried out but additional analysis could be conducted. By the same token, reporting and dissemination could be further improved. L (Little) means little has been done and N (None) means none has been done. N/A means "Not applicable". As shown in

the table(9), many cells are labeled as "N" indicating none has been carried out. The conclusions are drawn from reviewing the reports of the student assessments and other relevant documents such as short briefs and results pamphlets.

Table(9) **Information on the level of analyses, reporting, and dissemination of the student assessment data.**

	MoE-DET National Test (Math, Science, Arabic, English)	NCHRD NAfKE (Math, Science, Arabic)	MoE-DET Tawjihii (Math, Arabic, not many others)	NCHRD TIMSS (Math, Science)	NCHRD PISA (Math, Science, Reading)	Schools School-based Assessment	MoE-USAID Early Grade Assessment (Arabic, Math)
Analysis							
Rank-ordering, Avg. and %	F	F	F	F	F	P	F
IRT or Reliability Test	P	F	P	F	F	N/A	P
Cross-unit Comparative Analysis	P	P	P	F	F	N/A	F
Multi-year Trend Analysis	L	F	L	P	P	N/A	N/A
Criterion-based comparative analysis	P	P	N	F	F	N/A	P
Explanatory Modelling	N	P	N	F	F	N/A	P
Reporting							
School Report (card)	P	N/A	P	N/A	N/A	N	N
Field Directorate Report (card)	F	N/A	P	N/A	N/A	L	P
National Education Report (card)	F	P	P	P	P	N	P
Data made available for others	L	P	P	F	F	P	F
Dissemination							
Schools	F	N/A	F	N/A	N/A	N/A	P
Field Directorate	F	N/A	P	N/A	N/A	N/A	P
MoE's General Directorates	F	F	P	F	F	N/A	N
International	N/A	N/A	N/A	F	F	N/A	P

The table above indicates, for example, that NT results presented in annual reports tell us that only simple descriptive data analyses showing averages or percentages by various categories have been conducted. Even the descriptive results presented in the reports tend to be long and repetitive. Policy argument or implication for explaining the level of student NT performance is not discussed in these annual reports. No policy briefs are written as a result of the annual analysis of the NT data or report. This could potentially be an area of improvement. TIMSS and PISA have been managed and administered by NCHRD. The data analysis and reporting of TIMSS and PISA are more extensive in scope and on technical grounds than that of NT. However, in terms of advance level of analysis, comparable to the international norm of other participants in TIMSS and PISA, education policy guided data analysis could be conducted. Advanced modeling techniques could be applied to further explain why students perform differently even though they were in the same class, taught by the same teacher, or managed by the same principal. We believe that there is further need at NCHRD for developing policy-guided data analysis so that evidence-based policy implications could be suggested in the report.

IV.3. Overall Synthesis of "Student Assessments" in Jordan

Indisputably, the essence of student assessment is to produce timely, valid and reliable outcome measures of schooling process for a specified period of time and at a specific stage of student learning capacity, and then to let the results be usable and useful through analytics for policy development aiming at improving the quality of education and advancing the continuous learning journey. The analytics may require us to relate the outcome measures to varying input and process of schooling

contexts. But most importantly, any analytical or statistical results of student assessments must be examined within comparative analysis frameworks to generate new "policy expectations."[39]

IV.3.1 Analytics

Based on the analysis of annual reports of the assessments, we concluded that there is a lack of data analytics due to several reasons as we pointed out earlier, including: 1) lack of analytical skills, 2) lack of policy relevant demand, 3) lack of data sharing, and 4) lack of relevant data integration, and/or lack of time or job descriptions for conducting necessary analysis. In addition, we found that there have been limited findings from these assessments, which would provide 1) historical comparison (trend analysis or "self-comparisons" over time), 2) between-unit and within-unit comparison of averages and deviations (e.g. between and within schools or directorates variances), and 3) comparison against the pre-determined standards (e.g. set targets or yardstick). With all the assessment data available, we remain limited in integrating data from multiple years, sources and ability levels. These are serious shortfalls toward achieving the objectives of the student assessments. There is no doubt that any perfect design of the student assessment, if it is without a strong analytics regimen in place as part of the assessment system, is just as undesirable as any other poorly designed system. The key lies in the analytics and how the results become usable and useful. This is the area that Jordan's MOE should make tangible efforts to significantly improve.

IV.3.2. "Redundancy"

Jordan does not have excessive or over-burdening student assessment systems in country in comparison with other comparable countries. Education key stakeholders in the MoE and the field directorates and schools shared similar perceptions. In terms of spacing and time intervals of student assessments between grades, subject domains, test staking level, Jordan is quite balanced. Jordan has only one census-based assessment (the NT) for Grades 4, 8, and 10 in math, science, Arabic and English (core subjects). Other tests such as NAfKE, TIMSS, PISA, and EGRA are sampled tests. This mixture of student assessments is not only necessary but critical to providing national level policy makers with essential performance information to monitor how students are learning in schools at all three levels of the education system—and to understand what policy actions must be made for the quality enhancement.

TIMSS, PISA and NAfKE are all sample-based testing tools for diverse purposes. For example, TIMSS and PISA are highly recognized international standardized tests for measuring common knowledge and skills of children of various grades across different nations. NAfKE is developed to assess how students perform in learning the new curriculum developed on knowledge economy skills. None of these tests any form overlap with the domestic NT assessment. They have different purposes, measures of domains, intended comparability and utilization. One most important aspect of Jordan's participation in TIMSS and PISA is to put Jordan on the world map of assessed knowledge and skills among school children. At the macro level, it is critically important for Jordan to be well informed of how Jordanian children are performing in learning in comparison with children from other countries.

[39] Policy expectation means in this context new benchmarking or criteria for evaluating the assessment results.

"The number of tests is good and suitable and the questions that are asked are taken directly from the curriculum. It is not necessary to have additional tests in the future because [students] have enough tests." (Focus group notes, students, Irbid Town Prep (G/S3) for Girls, Irbid).

"Students should be tested regularly. Tests are important for the following reasons: tests help to identify individual differences and teachers can know the level of their students. Written tests are more suitable then oral tests. So testing students is important because it encourages them to get high marks." (Focus group notes, parents, Irbid Modern School, Co-Ed, Irbid).

"I don't have a problem with how many tests our students take per year, yet there is a lack in motivation and knowledge about these tests. I want every student to know about the importance of the tests they take. In some countries incentives were given to students to take TIMSS, for example." (Focus group notes, faculty, University of Jordan, Amman, February 4, 2014).

VI.3.3. "Stakes"

A test, particularly of a national scale, often has "stakes" attached to it. The stakes could be high or low for students, teachers, or schools regardless whether it is designed by policy intention or unintentionally. Usually, the stated purpose and/or use of assessment test would determine the level of stakes. By definition, a high-stakes test is "any test used to make important decisions about students, educators, schools, or districts, most commonly for the purpose of accountability—i.e., the attempt by national or local government agencies and school administrators to ensure that students are enrolled in effective schools and being taught by effective teachers. In general, high stakes means that test scores are used to determine punishments (such as sanctions, penalties, funding reductions, negative publicity), accolades (awards, public celebration, positive publicity), advancement (grade promotion or graduation for students), or compensation (salary increases or bonuses for administrators and teachers)[40]" In Jordan, both high and low stakes student assessments exist.

Although there has been no stated purpose of accountability for any test in Jordan, test scores are used to determine punishments, accolades, or graduation certification. For example, NT is known for grade promotion or repetition even though it is non-standardized student assessment, and Tawjihii is known for certification and admission to college or university. Without any "consequences" associated with the result by test participants, stakes can't be high for individual participants. A higher order importance such as monitoring system level performance and evaluation will be diminished if there is no accountability or follow up actions based on the results of the assessment. In principle, the level of stakes can be specifically managed or designed by policy intention.

To ascertain the stakes of the student assessments in Jordan, we interviewed students, teachers, school administrators, and other education stakeholders in the MoE and field directorates to get their perspective on how the importance they attribute to each assessment (see Table 10). Based on their perceptions of these tools, we estimated points of "stakes" for each assessment system in Jordan. For

[40] http://www.greatschoolspartnership.org

the highest stakes, we would give 10 points to an assessment system; for the lowest stakes, we would give only 1 point.

Table(10): Stakes' perspectives on how the importance they attribute to each assessment

"Stakes" Attached to the Assessments in Jordan				
	Students	**Teachers**	**Schools**	**Avg.**
NT (by MoE)	3	4	5	4.67
NAfKE (by NCHRD)	1	1	1	1.00
School Assessment (MoE and all schools)	6	6	7	6.33
Tawjihii (MoE, special unit)	10	9	8	9.00
TIMSS (NCHRD)	1	2	2	1.67
PISA (NCHRD)	1	2	2	1.67
EGRA (MoE in partnership with USAID)	2	3	2	2.33
Explanatory Note: The ranked scale of 1 through 10 is given to each assessment system in Jordan based on qualitative data on "knowledge" and "perceptions" from focus groups with students, teachers, and school administrators and other educational stakeholders. 1-point means "I have never heard of it" or "not important to anyone". Some remarks often expressed as "I don't care and never prepare for it" or "never hear or think about it"; 10-point means "everyone knows it" or "everyone has to prepare for it". Some remarks expressed as "it is the most important test in my student career" or "my future depends on it" or "as a teacher, I have to teach to the test"				

From the table above, it is clear that Tawjihii has the highest stake for students. Students in secondary schools are required to take Tawjihii before they can graduate with diploma and get admitted to college. It is used by all universities as the criteria for admission. It is not uncommon for students to spend, in some occasions, several years in private tutoring or supplementary classes to better prepare for the Tawjihii test. In our focus group discussion with students from Grade 4, we find almost all of them know about the Tawjihii and reported how important that will be for their schooling career in their future.

Although teachers and school administrators are not accountable if their students do not perform well in Tawjihii towards the end of the secondary schools, they all desire to produce higher number of students who do well in it. It is a teacher pride or school-wide pride that creates a high stake for them.

> "In 1964, Tawjihi was used as the system to evaluate students and send them to respective majors. It is not a working system, it has failed. The change is easy yet there is need for a political decision so that it can be carried out. It's not a matter of the curriculum or the semester system, etc. All the decisions are not based on research. We need political determination." (Focus group notes, faculty, University of Jordan, Amman, February 4, 2014).

NAfKE, compared to the other assessments, has the lowest stakes. Although the test has an important stated purpose of measuring knowledge economy skills and a critical mission of monitoring and evaluating the bottom line results of the ERfKE reform program, it does not have any consequence or accountability attached to the performance level for students or schools. Thus there have been no stakes in NAfKE. Although it is a sample-based assessment, many students—and even teachers— from

those participating schools do not know they participated in the recent past. We often heard from focus groups they never heard of NAfKE or ERfKE at all. There has been no single student or teacher or principal who considers NAfKE assessment as a high stake. In fact, almost none of them could identify any importance or use of it, and no one prepares for the test. This may in fact present a truism in reflecting the learning performance reality in Jordan. NAfKE is just like TIMSS and PISA tests which have low stakes but may reflect a real reality of student performance in Jordan.

Clearly, the TIMSS and PISA assessments are regarded as international tests which do not have much local stakes. Schools are randomly selected to participate based on the sampling rules proposed by IEA or OECD agencies that are in charge. The select schools don't teach to or study for the test and students neither prepare for these assessments, nor do they frequently even remember having taken them.

V. Recommendations towards More Integrated Systems of Student Assessment in Jordan

It is well recognized that any student learning assessment tool has limitations, but is necessary to assess students' learning performance for a specific purpose of improving teaching and learning, as well as monitoring and improving the quality of education. Multiple assessment tools for various subjects, domains, and different grade levels, particularly systemically used throughout the schooling process over time and for the recognized cognitive[41] and psychological development stages among students[42] can become an integrated and effective educational assessment system. Careful design, systemic development and upgrade, higher-order analytics, and reliable administration are all critical. While we applaud Jordan's established student assessment systems as we already described in this report, we find that Jordan is facing more than ever the growing challenges to enhance the quality, relevancy, analytics, usefulness, and trustworthy of the existing assessment systems. Some relevant questions have already been surfaced for some time but have not been systematically addressed such as: 1) How to integrate various assessment tools into strategically organized and effective educational assessment system? 2) How to design and upgrade an assessment tool within the educational assessment system that is locally curriculum relevant, 21st century skills germane, scientifically rigorous, and supportive of the teaching

[41] Jean Piaget, a Swiss theorist and psychologist, believed that "intellectual development takes place through a series of stages, which he described in his theory on cognitive development. Each stage consists of steps the child must master before moving to the next step. He believed that these stages are not separate from one another, but rather that each stage builds on the previous one in a continuous learning process. He proposed four stages: *sensorimotor*, *pre-operational*, *concrete operational*, and *formal operational*. Though he did not believe these stages occurred at any given age, many studies have determined when these cognitive abilities should take place." (Reese-Weber, Lisa Bohlin, Cheryl Cisero Durwin, Marla.*Edpsych : modules* (2nd ed. ed.). New York: McGraw-Hill Humanities/Social Sciences/Languages. pp. 30–132.)

[42] Erik Erikson, German-born American psychologist, developed eight stages of psychosocial development. "Stage one is trust versus mistrust, which occurs during infancy. Stage two is autonomy versus shame and doubt, which occurs during early childhood. Stage three is initiative versus guilt, which occurs during play age. Stage four is industry versus inferiority, which occurs during school age. Stage five is identity versus identity diffusion, which occurs during adolescence. Stage six is intimacy versus isolation which occurs during young adulthood. Stage seven is generativity versus self-absorption which occurs during adulthood. Lastly, stage eight is integrity versus despair, which occurs in old age." (http://en.wikipedia.org/wiki/Erikson%27s_stages_of_psychosocial_development)

profession? 3) How to develop a national "item bank" and reliable items for all levels and subjects in Jordan to meet the dynamic and growing demands of all stakeholders to evaluate student learning and school performance? 4) How to provide diverse types of succinct and useful assessment information or results to all stakeholders? 5) How to gain the public trust when the student assessment data is released? Although these challenges can't be all resolved quickly by this study report, the report presents an inaugural recommendation for a significant upgrade and enhancement of the student assessment systems in Jordan. We also must point it out that Jordan is not alone facing these challenges the world is also seeking answers or solutions. For this reason, we recommend a strategic vision and thinking for a more integrated system of student assessments in Jordan.

Sorting students into high and low performance categories based on the assessment results, for example, may be useful for targeting resources and providing extra efforts for the performance needy, but may also result in a negative consequence if actions are taken as sanctions against poor performing students or schools without extra support and assistance (Senge, 2000; Baker 2011). While we understand that a standardized testing system, particularly multiple-choice testing, may not align well with the goal of developing a set of new competencies to succeed in the global economy and society (Harvard Education Letter 2013)[43], we strongly believe that a well-integrated student assessment system that includes various forms and designs of assessment tools for the purpose of improving equitable learning and achievement is imperative in Jordan. The integrated student assessment system, which should include the standardized and non-standardized assessment tools, should aim to foster the development of the students' ability to analyze, synthesize and make inferences from data and facts. Today, the question is not whether Jordan should have assessments or not, or which kind of the assessment to choose from, but how to develop the integrated smart assessments, implement with the appropriate strategies, conduct the right analytics, and act on the results with positive and progressive solutions. Jordan needs to do it right and do it well.

Jordan's education is facing a critical and historical moment of globalization, technology advance and geo-political changes. Since 2003, the major education reform program, ERfKE, has led to significant changes in curriculum, school environment, management, teaching and learning, as well as instructional technology and assessments. All these changes are intended to improve student learning for the 21st century knowledge economy. Undoubtedly, the changes will continue and as the Greek philosopher Heraclitus said "change is the only constant;" this seems to be especially true in the modern era. We believe that the curriculum will continue to change (dynamics of curriculum), pedagogical practice in classroom will continue to change (such as the concept of "flipped class") and all things schooling will continue to change. So too, should learning assessment.

We understand the student assessments in a holistic view and therefore approach their development and improvement in a holistic manner. Not only should Jordan capitalize on the development work already done under the ERfKE program in the area of student assessments but it must also utilize the existing local capacity to advance and upgrade the assessment systems with a technical assistance

[43] Published bi-monthly by Harvard University, Graduate School of Education, Harvard Education Letter releases the latest research in the field of education by professors, scholars, researchers and education analysts.

support. The integrated framework emphasizes on "the after-test analytics, utilization and policy actions" beyond the assessment design and development. It is well recognized that the work of student assessments is an on-going development process, with new measures and test items added and the old and outdated withdrawn. Jordan must have strong and sustainable technical forces and institutions to continuously develop and upgrade the assessment systems. It is educators' responsibility to meet the challenges of the dynamic education world.

V.1. Definition

We define the more integrated systems of student assessments as more comprehensive, state-of-the-art, internationally comparable and local curriculum-based but better balanced, complementary, value-added as well as intra-connected systems. The next systems in Jordan, although they are not new, must also be more trustworthy, relevant, congruent, and "stakes" controlled. We will explain this in more detail later in the report. This approach requires a sensible and significant "update and upgrade" of the existing student assessment systems in Jordan. The systems must also be supported by the diverse institutional capacities that are already established in recent years and the systems must be able to provide timely, reliable, and useful assessment data to all the stakeholders in a transparent and collaborative manner. The proposed integrated systems in this report include our recommendations for going forward based on the results of this Mapping study.

It is noted that we were not clear about precisely how integrated systems would look when the Mapping Study was initially launched. Terms such as consolidation, test replacement, getting rid of redundancy or burden, and shared responsibilities were pre-conceived to explain what this study could address or focus. Once the study was undergoing and analytics efforts began to reveal findings, it became clear to us that much improvement is needed over the current assessment systems and beyond our initial scope of the study. Then the idea of "adjustments and integration" becomes more and more mature and feasible, which is necessary to meet the challenges for producing globally competent, technologically savvy, and the 21st century skills equipped graduates of all levels in Jordan in a holistic way.

V.2. Necessary Adjustments and Integration

Given the fact that most countries have student assessments that serve at least three out of all four major purposes (i.e. 1- learning improvement, 2- graduation, 3- punishing or incentivize performance, 4- accountability), we believe that Jordan is no exception. The current four domestic assessment systems in Jordan serve three purposes (all mentioned above but accountability). By our evaluation, the systems are appropriately structured, at least by design or on paper. For example, the NT is designed for the monitoring and improvement of the overall student learning of the new curriculum for the knowledge economy skills; school assessment is used to sort the students into the good and poor performing learner categories; Tawjihii is clearly used to certify the high school graduation and award a college entrance; and NAfKE is used to measure the bottom-line outcome of ERfKE implementation and what program factors might explain the changes in that outcome. These systems are all necessary and have been in existence for years, but they are not as well integrated as they should be.

In addition, Jordan has been one of the few but early participants in the Middle East region in TIMSS and PISA, and recently began to participate in the USAID supported EGRA assessment to fill the void of early grade assessment. We strongly recommend that Jordan continue to participate in the international tests and the EGRA assessment but make the proposed changes below to cultivate more value-added and integrated for the domestic "consumption", which will be covered in greater depth later in the report.

Three Structural Adjustments and Integration

1) Increase the effort to assess all students learning performance in Grades 4, 8, and 10 annually and raise the stakes of the NT assessment system.
2) Convert NAfKE to NAfKE-JOR and expand its plan to assess students in Grades 3, 6 and 9 as a sample test; and raise the stakes.
3) Continue to participate in TIMSS and PISA but add Grade 4 TIMSS, and use EGRA for early grade assessments.

To illustrate the proposed adjustments, the following table (11), lists all assessments and color-coded (red) them accordingly.

		Grades to be tested											
		1	2	3	4	5	6	7	8	9	10	11	12
Census Assessments	NT (annualized)				x				x		x		
	Tawjihi												x
Sample Assessments	NAfKE_JoR			x		x	x			x		x	
	TIMSS				x				x				
	PISA									15-yr. old			
Sample Assessment	EGRA		x	x									
MoE's ongoing assessment	School Assessment (SA)	All grades											

NT – National Test, managed and administered by the MoE, must have a sizable item bank of relevant domains, subjects, and grade levels. To raise NT stakes, results could be considered as part of school assessment for Grades 4, 8, and 10 students.
Tawjihi – General Secondary Certificate Examination in Jordan, currently managed and administered by the MoE
NAfKE_ JoR – National Assessment for Knowledge Economy Skills _Jordan, Reading and Math for grade 3 & reading, math and science for grades 5 and 9. It takes place every two years.
TIMSS – Trends in International Mathematics and Science Study (international). Jordan has participated in TIMSS for the past 4 cycles since 1999, but only in grade 8, not in grade 4. NCHRD manages and administers TIMSS
PISA – Program for International Student Assessment (international). Jordan has participated in PISA for the past 3 cycles since 2006. NCHRD manages and administers PISA
SA - School assessment is an on-going throughout an academic year, student's cumulated composite score may be from 1) subject learning performance (quizzes and tests), 2) discipline (behavior), 2) social responsibilities (peer support, community duties, and school tasks)

Note: Red color **X** indicates a new addition or major change, green color box means "cancelled", and dark color X means no change.

V.3.1. NT System Adjustments, Item Banking, and Raising the "Stakes"

The MOE should continue to manage and administer the NT system, but cover the core subjects of all 3 grades, (4, 8 and 10) annually. Making it an annual routine is critical so that no students should skip or miss NT test as they reach each of these grade levels, 4, 8 and 10. Currently, students of many cohorts in Jordan skip or miss the NT test(s) throughout the schooling system as summarized in the following table. This is largely due to the current "alternation" process (NT test for one grade only in a given year) in Jordan as described earlier. For example, in year 2001, if students entered school that year in grade 1, then they would have skipped every NT test year for grades, 4, 8, and 10. So would the cohorts of year 2004, 2007, 2010, and 2013 and so on. Student cohort of 2002 in Jordan would have skipped NT tests for grades 4 and 8 in 2005 and 2009 respectively, only caught grade 10 NT test in 2011.

	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11	Grade 12
2000			skipped	Y								
2001	skipped		skipped	skipped				Y				
2002	skipped	skipped		skipped	skipped					Y		
2003		skipped	skipped	Y	skipped	skipped						
2004	skipped		skipped	skipped		skipped	skipped	Y				
2005	skipped	skipped		skipped	skipped		skipped	skipped		Y		
2006		skipped	skipped	Y	skipped	skipped		skipped	skipped			
2007			skipped	skipped		skipped	skipped	Y	skipped	skipped		
2008				skipped	skipped		skipped	skipped		Y	skipped	
2009				Y	skipped	skipped		skipped	skipped			skipped
2010						skipped	skipped	Y	skipped	skipped		
2011							skipped	skipped		Y	skipped	
2012				Y				skipped	skipped			skipped
2013								Y	skipped	skipped		
2014										Y	skipped	
2015				Y								skipped

The current NT system does not provide a complete and comprehensive assessment of all students for different levels of education. Annual assessment of all 3 grades is necessary and critical for Jordan to monitor the quality of its complete education system. No other assessment in Jordan could substitute this scope and scale of the purpose given its anchoring capacity. The NT system must also be the curriculum-based and nationally standardized assessment system. If so, it could truly serve the purpose of improving teaching and learning at school, directorate and national levels. Specific use of the NT results should be further consulted with policy makers in the MoE and thoughtfully considered for evaluating teacher performance as well as school value-added education for learning purpose.

NT system, once well set as proposed, should enable three comparability and analyses that the MoE currently lacks, including: 1) longitudinal analysis (including tracking students), 2) cross-unit analysis, and 3) comparative analysis against the national standards or expectations. Each one of the three analyses may be applied independently from or in conjunction with the others. For example, the cross-unit analysis with within- and between-variance components analysis (illustrated earlier) could be conducted through the longitudinal analysis method to examine the trends (tracking students, teachers, and/or schools). This allows a closer and more frequent monitoring and tracking of the cohort performances from year to year in the learning achievement (Grades 4, 8, and 10) and how the changes (improvement or decline) in learning achievement are and what may explain the changes. Simple information of

student characteristics should also be collected including student gender, age, and ethnicity/refugee status.

NT system must allow the MoE to frequently monitor the national "pulse rate" of the learning performance and the school quality and develop action plans accordingly for the improvement. It would allow the policy makers and researchers in Jordan to track individual students longitudinally and design a unique tracer studies to identify what teaching and learning interventions used in earlier grades may impact the student learning in later grades. It would, for the first time, let the MoE conduct the value-added analysis of the school effect.[44] It would allow the MoE to annually produce "nation's report card" and individual (single pager) school report card for all the schools on the educational progress in Jordan or in schools.

Undoubtedly, the NT system must have a sizable item bank for relevant domains, subjects and grade levels. Multiple categories of sub-domains could be envisioned for the bank and specific items under each subject or grade level could be developed and/or borrowed (or purchased) from credible and reliable sources internationally. For example, there are many items already developed to measure critical thinking skills, problem-solving skills, and synthesis skills in many test centers around the world for different grades and subjects. Jordan can surely match them to its own curriculum needs and the ERfKE requirement of the 21st century skills in addition to its own existing items embedded in the MOE and NCHRD. The Item bank development is known as an on-going development process which requires a significant national effort to manage, coordinate and maintain. Jordan should no longer wait and the MoE/DET should have a few NT item bankers.

The NT must also raise its stakes since it is a curriculum-based assessment so that students and teachers take them more seriously for teaching and learning purpose. In order to do so, the NT assessment results could be considered as a part of the school assessment for Grades 4, 8, and 10 students. For example, instead of having teachers develop their own final exams locally for these grades, the NT assessment results could be used as part of the requirement, (for example, up to 40% stake) in the final school assessment report. This will increase the perceived "stake" by students and teachers. It must be noted that the preparation for a test is a learning process, even in the current drive for the knowledge economy skill, particularly if the test items are measures of the critical thinking, problem solving and synthesis skills. We also strongly believe that Jordan's national report card and individual school report cards, developed properly, would raise the NT stakes. Revealing the NT performance results to all stakeholders through the comparative lens and the report card mechanism would contribute to greater transparency in developing the system-wide culture of data for educational decisions. It would surely bring about the higher stakes that the NT deserves.

[44] The value-added school effect study requires tracking students and conducting several types of "learning gains" as the standardized outcome of the achievement and then singling out school net contribution (value-added) controlling for student, household and other social but outside school characteristics or factors.

V.3.2. National Assessment for Knowledge Economy – Jordan (NAfKE-JOR)

NAfKE-JOR is a newly proposed assessment system that is inspired by EGRA and PIRLs (http://timssandpirls.bc.edu/) and informed by the NAfKE experience but complementary to the NT system. It makes a perfect sense that Jordan gives priority to literacy and numeracy earlier in students' school careers. The assessment takes place in grades 3, 6 and 9. This would avoid unnecessary duplication with NT grades. Grade 3 will be the earliest grade for student learning assessment at a national level in Jordan.[45] This assessment must be sample-based, research-oriented, and well-designed. The domains and assessment items could be unique and more progressive. Many of the assessment domains and items may be expanded beyond those used by the NT assessment. For example, creative writing, analytical synthesis, numeric estimation, mathematical problem solving, partial credits through Item Response Theory (IRT), etc. could be all designed in the NAfKE-JOR.

The NAfKE-JOR, although it should be considered as a national comprehensive assessment, should not be an annual assessment but an assessment that could be recurrent every three years. In addition, the spacing among the grades is also suggested to be three grades in sync with the testing years. This once again allows a unique design of tracking students and grade cohorts at the same time, which permits a great opportunity for the value-added analytics framework.

The NAfKE-JOR should be a nationally representative sample-based assessment that focuses on reading and math literacy in Grades 3, 6 and 9. The purpose would be to provide an "early warning and learning" information for improving basic causal core (reading and math) of learning in education. This should be managed by external NCHRD. The framework for this assessment can be adopted from the EGRA and the NAfKE frameworks. Domains and the associated measures in early grade could cover from phonemic awareness and decoding in reading to counting and shape recognition in math. In higher grades, domains and measures may cover reading fluency and comprehension in reading and writing, and algebra, problem solving, and statistics in math. Details in this newly proposed system could be further developed once the MoE decides in principle to adopt the recommendation.

The NAfKE-JOR should provide additional and external information to educators how students are learning in the fundamental core of the schooling and the data should inform a higher order research and analysis. For example, the results from the NAfKE-JOR must help identify the three types of students and schools in the view of the learning gain achievement (calculated by using multiple years of the tests) over time: 1) the "status quo" in performance maintained, 2) the performance deteriorated, and 3) performance significantly improved.

Although they can be identified fairly easily with the right assessment design, the following questions may be difficult but can be answered: 1) To what extent are the three types of students all within the same schools? 2) Do many schools in Jordan categorically belong to one of the types? 3) What could be

[45] Assessment (tests or exams), earlier than grade 3, may cause larger than expected measurement errors given an immature stage of children development. However, grade 3 is early enough to detect any potential early sign of needs for teaching and learning related problems at all levels.

school factors that variably contribute to the scenarios mentioned above? 4) What could schools do in terms of developing policy actions to address the learning problems if identified?

We believe that the NAfKE-JOR system, with the proper design, serves a national need of conducting a longitudinal research to answer these types of education sector policy questions. Again, the NAfKE-JOR must track Grade 3 students and assess the same students when they reach Grade 6 after 3 years of the assessment cycle. The same could be applied for the Grade 6 students going into Grade 9. This will add value to the sample based student assessment with this design. The NAfKE-JOR is not a redundant effort but necessary research-oriented necessary. It should not have a high-stake as NT for students or teachers, but valuable stake for researchers and education policy makers in Jordan.

The real value of the NAfKE-JOR lies in the development of a new outcome measure that is different from the traditional multiple choice type assessment. It should be truly targeting the 21st century knowledge and skills to be obtained by students in Jordan. It may be even possible in the future that the NAfKE-JOR takes a lead in measuring the "five minds" for the future (Gardner, 2006)[46] in Jordan and the new social and "global competences" under the NAfKE-JOR's research-oriented assessment framework.

We also recommend that the NAfKE-JOR system develop its own item bank. Given the small scale and less frequent nature of the assessment, the number of items may not be as large as that of the NT system. The earlier recommendation regarding the NT system item bank development could be applied here too for the NAfKE-JOR's item bank.

V.3.3. Participate in Grade 4 TIMSS

Under the new integrated student assessment systems in Jordan, NT student assessment only covers grade 4 students for the purpose of monitoring system level student performance. Participating in TIMSS grade 4 level (Jordan already but only participates in grade 8, and most of participating countries in the world participate in both grades 4 and 8) is necessary and critical to see how grade 4 students in Jordan perform in the context of global competitiveness. NT assessment, although valid within Jordan context, is not designed for the comparative context. We therefore strongly recommend that Jordan participates in grade 4 TIMSS. By participating in Grade 4 TIMSS study, Jordan will be able to design a long term strategy to track students over time from Grade 4 to Grade 8 through the two cycles of TIMSS (four years). Participating in international student assessment requires a fee of 40,000 US dollars, but the benefit from the participation could be invaluable, particularly if Jordan well utilizes its results for benchmarking purpose and education system improvement.

V.4. Other Important Recommendations

School-level Assessment

Schools must be able to assess students frequently in academic or learning performance, school discipline, and social responsibilities in a learning environment. The three areas of student performance outcome measures in schools are highly correlated according to the MoE staff. Students must be

[46] "Five Minds for the Future", according to Harvard Professor, Howard Gardner, are: 1) the discipline mind, 2) the synthesizing mind, 3) the creating mind, 4) the respectful mind, and 5) the ethical mind.

encouraged to do all well in schools. Many high performing school systems around the world are also emphasizing all-round performance measures for students (e.g. Finland, and the state of Massachusetts in the US). The school assessment framework must be guided by the MoE and field directorates and developed by schools. Final student results should be collected by the field directorates for aggregated analysis across schools within each directorate. The key is that schools must make the school assessment criteria transparent to parents, students, teachers and other stakeholders. The idea is to let all stakeholders be informed of what and how students are continuously assessed throughout school year in schools. The stake that is raised by informing all stakeholders is effective and useful. There are only benefits to learning and teaching communities. It is admitted that there could be more "pressure" for students but there is no drawback in this case.

For academic performance, teachers must be at the forefront of this process. As recommended by Harvard educators who developed 8-step "Datawise" program for teachers, rapid assessment by and quick feedback to teachers is not only necessary for 21st century teaching and learning, but also a new required competency for any teaching profession[47]. If teachers have the ability to test students but do not have data literacy, testing will not be useful for teachers or students for an improvement purpose.

For school discipline, students studying in schools must learn to how to work and learn in a group or with peers. School discipline is the system of rules, punishments, and behavioral strategies appropriate to the regulation of children or adolescents and the maintenance of order in schools. Its aim is to control the students' actions and behavior. Behavior among peers in a class must be productive and conducive for teaching and learning so that learning time in class is maximized. Mutual trust and respect for each other and being kind towards others must be promoted in school. If this gets well developed and rigorously implemented as part of the school-level ongoing assessment, student learning performance may also be improved. To paraphrase former U.S. President Bill Clinton, one of three problems in the US education system is the student discipline problem in class or school.

New Strengthened Capacity in Student Assessments in Jordan

Both the MoE and NCHRD may need to strengthen their institutional and technical capacity to ensure both kinds of utilization of student assessments (data and results) in addition to the capacity of the design and management of the test items bank. Jordan should avoid the scenario of "only investing in a good design of student assessment system without capitalizing on the utilizations of the system development." As Hua (2003) said, "any (well designed) data system, if without a routine set of analytics put in place, without an institutional policy of data sharing, without specific and policy relevant information products produced regularly, would be an useless system that only consumes a large sum of resources."

In terms of data analysis, an overall analytical framework must be well understood and pursued and advanced. While international corporations use the "big data" for meeting their customers' or clients' needs and satisfaction, the education sector must use the existing data including assessment data to address the needs of learning and teaching. Within this framework, any analysts must first envision

[47] http://www.gse.harvard.edu/news-impact/2012/01/the-data-wise-process

three comparative contexts under which all statistical analyses must be conducted and applied. They are 1) longitudinal or historical context (also considered as trend analysis or self-comparison); 2) cross-unit or between-unit comparative context; and 3) comparative context against the standard or the expected (stated quantitative and qualitative targets or objectives). Although there are distinct features among these comparative analyses, they are not by any ways mutually exclusive of each other. In fact, we can often apply all three of them for a single but comprehensive analytical undertaking.

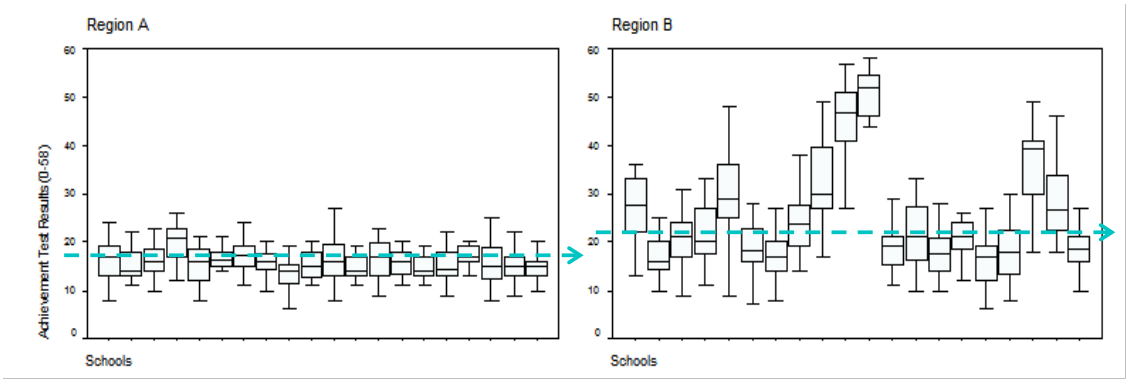
Longitudinal Comparison

It is critical that the MOE plan any student assessment with a long term vision. Not only does this require a good design of the student assessment (items, samples, and administering process), but also it requires a "comparability" across the time. For example, Grade 4 test results from two separate years are bound to be different, but how the MoE can decide to create anchor items in order to equate the two tests for the reliable comparability requires a strategic planning in the item design and item selection process.

The education research community in Jordan—in tracking individual students over time about their learning outcomes as well as learning characteristics—may help explain the unexplainable phenomenon. For example, from this type of design, we may be able to answer why girls and boys in Jordan perform equally well in achievement results in early grades but very differently as they move up to higher grades. Boys underperform girls significantly in every subject and grade for the rest of the education career after Grade 4. Longitudinal comparison is not only an important statistical method but also a strategic thinking in terms of detecting and identifying changes (progress or regress) over time and over "comparable elements." In education, this includes but is not limited to the following implications: a) tracking individuals' learning achievement over time and over similar criteria, b) tracking schools' performance over time and over different cohorts of students, c) tracking national trends over time and across similar performance measures. Analysts may also examine gaps or differences (in gender, among ethnic/migration groups, between rural and urban, etc.) and other variances between and within schools or directorates, as well as statistical relationships between and among outcome and explanatory variables. All of these require a good and strategic design and planning.

Cross-unit Comparison

Cross unit comparative analysis should go beyond analysis of simple averages. In Jordan, there has been a fair amount of cross-unit comparison in examining student assessments. Tables of average student assessment results by regions, school types, and field directorates are common practice in annual reports. While it is valid to present averages across comparison groups (with standard deviation statistics indicated), cross-unit comparisons in variances, trends, gaps, are almost non-existent in Jordan. This important aspect of systems analysis in evaluation of education performance is unnoticed or overlooked. It is likely that there is a lack of capacity in envisioning this type of technical analysis. Below is an illustration of school math achievement performance level (actual data from a sample country) to demonstrate two diversely different regions in the sample country that have slightly different regional averages (as two lines indicated) but very different "variances" across the schools and within schools.



This diagram tells that Region B has both large between-school variance and within school variance. Region B is of much more heterogeneous system than Region A. This "map" would provide policy analysts and policy makers with much needed information and further policy driven inquiries. What local policies or policy practices and/or conditions have caused the sharp differences in the two regional systems? What should be done next to improve the educational quality and equality? Is the Jordanian education system now being reformed heading towards a system more like Region A or Region B in 5 or 10 years?

Imagine we display 10 years of "system maps" like the one above in Jordan to examine the changes in variances over time, what could be that development trend and what could be a new policy dialogue in education development or reform? That would be a good case for combining the cross-unit comparison analysis with longitudinal comparison analysis.

Comparison against the Expected

Setting up tangible and indicative targets to measure an achievement of goals is required for conducting this type of analysis. The key is to define "the expected." Sometimes the expected could mean a range of quantitative results at a national level or at specific targets with variable resources or level of efforts or time frame. For example, to increase a national achievement score by one percentage point (on average) may require more resources than that to increase three percentage points in a few poor performance directorates. If both national (1 percentage point increase) and targeted directorates (3 percentages points) are provided with "the expected" targets to reach with a specified time frame, the comparison analysis against the expected could be very informative. Additionally, cheering over a statistically significant and positive finding may be too professionally naive. If the average reading level for Grade 2 students in Jordan is 5 words per minute, a 20% jump in results after one year of learning to 6 words per minute, while considered a statistically significant improvement, is sadly insufficient.

Setting up standards and policy targets is necessary to stimulate this type of a demand-driven comparison analysis. If the expected (e.g. benchmarking, targets, criteria, etc.) is not defined, there will be no meaningful analysis. As part of the integrated student assessment systems, all students, teachers, principals and other educational stakeholders in all schools for in all grades and subjects must be informed of and become knowledgeable about the national, directorate level and school level expectations. Currently, Jordan does not have the well-articulated national, directorate or school

expectations. These could come as a form of indicators or targets. As USAID's policy target (a specific policy expectation) suggested lately through EGRA initiative, "75% of all students in Grade 2 in Jordan should be able to reach the reading proficiency, 30 words correctly per minute." That is a well-articulated and specific expectation. Anything below that would be considered inadequate.

V.5. Moving Forward

According to the highly publicized McKinsey's report (2007), "***How the world's best-performing school systems come out on top***", three factors that contribute to all high-performing education systems in the world are: 1) getting the right people to become teachers, 2) developing them into effective instructors and, 3) ensuring that the system is able to deliver the best possible instruction for every child. Clearly, none of the key determinants mentioned is about student assessment system development. The student assessment system alone won't be the determining factor for a high-performing education system, but a well-designed and administered student assessment system, with results used effectively, acting in unison with other smart education policies, should become essential for monitoring the system and individual performance levels and informing policy actions for the improvement of learning and teaching. Without it, the catchword, "improvement" is simply an empty verbiage.

VI. References

Boudett, Kathryn; City, Elizabeth; and Murnane, Richard. "***Data Wise: A Step-by-Step Guide to Using Assessment Results to Improve Teaching and Learning***" Revised and Expanded Edition. Harvard Education Press. 2013

Eva L. Baker. "***Can We Fairly Measure the Quality of Education?***" CSE Technical Report 290. Center for Research on Evaluation, Standards and Student Testing, UCLA. 1988

EACEA P9 Eurydice "***National Testing of Pupils in Europe: Objectives, Organization and Use of Results***" Education, Audiovisual and Culture Executive Agency, European Commission. 2009

Hua, Haiyan and Jon Herstein. ***"Education Management Information System (EMIS): Integrated Data and Information Systems and Implications in Education Management"*** Comparative International Education Society. 2003

Ecclestone, Kathryn and John Pryor. "***Learning Careers' or 'Assessment Careers'? Impact of Assessment Systems n Learning***" British Educational Research Journal 2003.

NCHRD, Jordan. "***National Assessment for Knowledge Economy Study Report"*** *2007*

NCHRD, Jordan. *"**National Assessment for Knowledge Economy Study Report**" 2008*

McKinsey & Company Report "***How the World's Best Performing School Systems Come Out on Top***" 2007

Senge, Peter. *"**Schools That Learn**"* New York: Doubleday Publication. 2000