WHAT EXAMINATIONS TEST

Khaled Almadani
University of Dundee, Dundee, Scotland, UK
E-mail: K.A.A.Almadani@dundee.ac.uk

Norman Reid
Universities of Dundee and Glasgow, Scotland, UK
E-mail: dr_n@btinternet.com

Susan Rodrigues
University of Northumbria, Newcastle, England, UK
E-mail: Susan.rodrigues@ed.ac.uk

Abstract

Examinations play a dominant role at school level in all countries. They determine the future of candidates in many ways. However, what do national examinations actually assess? In a previous study in Bahrain, it was clear that students had reservations about the way they were being assessed.

This paper follows this up by examining examination data in all subjects undertaken for over 7000 school students in their final school examinations. The data were explored statistically, including the use of principal components analysis.

The findings reveal some interesting problems and, perhaps, the most pressing one was the finding that all examinations in all subjects were measuring one variable only. A look at examination papers showed this to be recall. The findings are related to previous findings and suggest a widespread problem in considering quality in secondary education in that schools will focus on the rewards given by national examination systems. Thus, assessment motivates the focus on memorisation.

Key words: *quality assurance, national assessment, memorization.*

Introduction

Almadani *et al.* (2011), in their survey of the perceptions of Bahraini school students as part of a study in quality assurance, found that one of the recurring features in their responses related to the place of memorisation. Many learners felt that the schools in Bahrain showed them what to memorise and they tended to see memorisation as dominant over understanding. Indeed, there was evidence that they wanted the freedom to think and to understand. This reveals an area where quality is being questioned.

Assessment probably holds the key to all this, for students and their teachers will focus on the skills which gain the greatest rewards in any assessment system. Thus, assessment motivates the focus on memorization and the way schools see assessment is, inevitably, strongly influenced by the national examination system. This study looks at some aspects of the national assessment system in Bahrain. The summative aspects of assessment will be the focus for this study.

Problem of Research

The aim of the study reported here is to look at the data obtained from the final year National Examinations in Bahrain, mainly from a statistical perspective, to see what insights can be gained about what is being measured and how the data are being interpreted. Although one specific country is considered, findings from other countries will also be considered in that it is likely that there are common patterns. Indeed, in an age of globalization, Little (1993: 13) notes that, '*...increasing evidence of the powerlessness of individual countries to stand outside the now international market of qualifications'*.

Research Focus

There are many descriptions of assessment (eg. Angelo, 1995; Palomba and Banta, 1999; ABET, 2003) and they all tend to focus on encouraging the quality of student learning. Thus, Angelo (1995: 13-14) see assessment as,

> *"An ongoing process aimed at understanding and improving student learning. It involves making our expectations explicit and public; setting appropriate criteria and standards for learning quality; systematically gathering, analyzing, and interpreting evidence to determine how well performance matches those expectations and standards; and using the resulting information to document, explain, and improve performance."*

In a more formal sense, assessment is the process of gathering, interpreting and using evidence to make judgements about the achievements of students in learning. The evidence can be gathered by looking at what students say, write or can do. Sometimes, useful assessment insights can be gained from teacher's reports or from observation of progress and performance in class. However, assessment tends to look at the '*product*' at the end of a piece of work, a module or a course. Nonetheless, it is equally possible to look at the experience of learning itself. Thus, all assessment involves reasoning from evidence, is imprecise, and is only an estimate of what a person can do (National Research Council, 2001). Indeed, everything we ever learn is stored in the brain and assessment means looking at what is stored and how it is used. This means that we have to deduce what is in the brain by observations of performance of the candidates and, inevitably, this is an imperfect skill.

There is a danger with all assessments. It tends to reward those skills which are easy to measure. It tends to neglect those skills which are difficult or impossible to measure. At a national assessment level, this has a '*backwash*' effect, encouraging schools to focus on what is to be measured, thus neglecting other, often more important skills, which are not to be measured. Thus, Broadfoot and Black (2001: 11) speak of aspects like, '*teaching to the test*', *'anxiety and low self-esteem', 'turning many students off formal learning forever'* as part of the '*backwash*' effect. Indeed, these effects arising from assessment were first identified by Messick (1989) where he describes them as '*wash-back*'.

One problem is that it is very difficult to be sure what is being measured for the learners have found all kinds of creative ways to obtain '*correct*' answers, many of which depend on recall. Certainty on validity is thus elusive. However, under sensible test conditions, consistency and reliability are usually assured although inter-marker reliability is a major issue in some subject areas (Hayward and Spencer, 2006).

The way assessment is used may be a real area of hindrance in learning. Outcomes can block future opportunities while assessment can generate knowledge reproduction and social reproduction. Indeed, testing can be used to control the curriculum and learning while tests can force a focus on test content, training students to the tests, practicing tests, and transmission styles of teaching (Broadfoot and Black, 2001). Of course, tests and examinations can generate student anxiety and loss of self-esteem. The students may only learn what is to be tested and there can often be a loss of student enjoyment in the process of learning. Thus, assessment can distort the whole process of teaching and learning.

There is a strong tendency in most countries only to trust the data from national examinations and to distrust teacher assessments. In fact, the evidence shows that teacher assessments are often very robust and, indeed, a combination of teacher assessment and national examinations is known to give the best data (Assessment Reform Group, 2006).

Examinations tend to give numerical data, often as percentages. It is too easy to see these in absolute terms. Thus, a score of 80% means that 80% is known, or a score of 50% means a '*pass*'. This, of course, can never be true. If an easy paper is set, the marks will be high; if the paper is more demanding, the marks will be lower.

A similar problem occurs when some examiners award very high marks, suggesting that the students in their subject are doing well. In simple terms, marks mean absolutely nothing. In Bahrain, everything is based on percentages and high scores are valued, it being thought that these indicate high levels of success. However, any examination is like a measuring instrument where the scale is not marked on it. Examinations merely place the candidates in an approximate order of merit.

Any national examination system must fulfil quite a number of purposes. For example, national awards may serve the following purposes:

- Recognising and rewarding what has been achieved over a sustained period of time;
- Reflecting accurately the agreed goals for the curriculum;
- Determining eligibility to enter higher education, and other career opportunities;
- Revealing specific skills and areas of achievement.

Examinations usually generate marks. Marks mean very little for they reflect the demand level of the specific examination paper. National examinations involve large numbers of candidates and it is possible to use the marks to place the students in a rough order of ability set against the criteria being measured in the particular examination. Thus, marks in any specific subject will generate a distribution that will be close to normal (Figure 1).

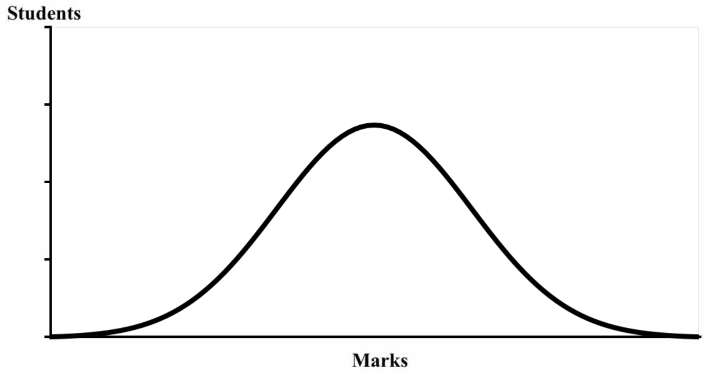

Figure 1: Normal Distribution.

The point on the curve which corresponds to the pass mark can only be determined by human judgement. Thus, the percentage gaining a pass is entirely dependent on the professional judgement of the examiners. Given a large number of candidates (perhaps about 1000 or more), then it is reasonable to expect approximately the same proportion to pass in successive years in any national examination. Thus, if 70% of the candidates passed last year in a specific subject, it is likely that a very similar proportion will have the similar abilities the following year and, therefore, deserve to pass.

The question is how to set the difficulty levels for examinations. It is important that the examination has high discrimination, making a clear separation between candidates of different abilities. This requires a high standard deviation. However, if the entire scale (from 0 to 100) is used and the mean is about 50 with this high standard deviation, then it means that half the candidates gain a mark below 50 and may feel '*failures*'.

Perhaps, a mean of nearer 60% might be more appropriate and a standard deviation between 10 and 12 might be more appropriate. This is illustrated in Figure 2.

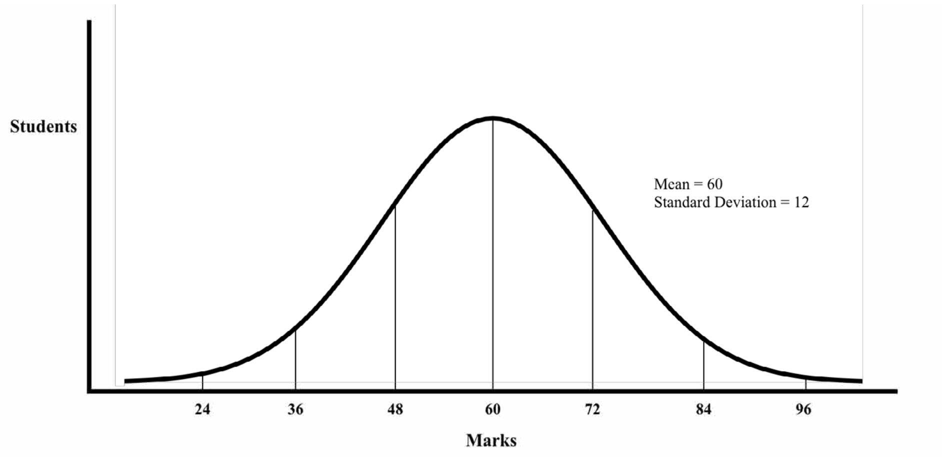

Figure 2: 'Ideal' National Examination Marks Spread.

In figure 2, only 16% of the candidates will score less than 48% while only 2.5% will gain more than 84%. This does not demotivate too many candidates while allowing the most able to show their abilities.

The above discussion has looked at examinations in any subject. The other aspect is what a *specific* examination in a *specific* subject actually measures. Of course, the content will vary from subject to subject. However, the skills will also vary. Thus, an examination in, say, a language might be expected to offer evidence of linguistic and literary skills and that will be very different from an examination in mathematics which might be expected to reflect logic-deductive skills, for example. In turn, the skills measured in a chemistry examination will be different from both mathematics and languages.

This can be explored by looking at inter-subject correlations. These will be positive (candidates tend to have all-round abilities) but the correlations should not be too high, reflecting the fact that the various subjects should be rewarding credit for *very different* kinds of skills.

There is another important dimension in looking at assessment. So often, assessment is seen simply as a set of examinations at the end of a course. Speedy *et al.* (2003) refer to the way assessment is a powerful force in supporting learning and, indeed, they see assessment in a restorative sense. However, the normal pattern is that assessment is seen as a filter, failing those who are deemed not to have performed well enough. Thus, assessment is a kind of '*certificate of failure*' for many learners. The idea that assessment can be used positively and affirmatively, rewarding achievement and suggesting ways forward for future learning is usually lost. However, if assessment is used to reward achievement and suggest ways forward for future learning, then it can be genuinely restorative.

Methodology of Research

General Background of Research

The examination data from all final year students in most subjects in Bahrain were collated, using spreadsheets. Three main curriculum pathways (science, literary, commerce) are considered separately, the uneven numbers reflect national student choices of pathways. The data include all Bahraini students in 2011 and all core subjects are included.

In all countries, national examinations reflect learner successes against specified criteria. Inevitably, they tend to be '*end-of-course'* assessments, involving large numbers of candidates, all sitting common examinations papers usually simultaneously. Examination outcomes have a powerful effect in determining the future career destinations of the learners as well as opening doors for higher levels of education. The examinations are designed to offer a measure of learner performance. In this, they are indicators of quality. However, they only offer a limited insight on quality for there are many outcomes from school education which are not open to formal national examinations. In this study, the aim is to look at the data obtained in one system of national examinations and explore what aspects of quality are being revealed.

Sample of Research

The sample is 7022 students in their final year at school in Bahrain. 7022 is the total number of students of this age in Bahrain following three of the four curriculum pathways. There are four types of schools at this level (age 16-18) in Bahrain (often on separate campuses, sometimes on one campus): scientific schools, literary schools, commerce schools and industrial-technical schools. The last group was not considered for it only caters for boys. The school students choose which curriculum pathway they wish to choose. Where the population density is high, they are housed in separate schools where each curriculum pathway has its own set curriculum. In more rural areas, the different curriculum pathways may be in one school but they are still separate pathways.

The numbers of student choosing these curriculum pathways varies enormously and, indeed, the gender distribution in these three curriculum pathways is very imbalanced. The samples here simply reflect the way the students have chosen their curriculum pathway.

Instrument and Procedures

The data were made available in the form of a spreadsheet by the Examination Authority in Bahrain. Principal Components Analysis was conducted using SPSS.

Data Analysis

The data came in the form of marks in all the subjects (as percentages). Some elective courses which draw very small numbers were not included. Means, standard deviations and histograms were obtained for each subject. Inter-subject correlations (Pearson) were considered. The data (for each of the three curriculum pathways) were examined by factor analyses, using principal components analysis with varimax rotation, to explore how many factors explained the inter-correlations. A minimum variance to be explained was set at 70%.

Each of the three curriculum pathways is now discussed, showing the inter-subject Pearson correlations, the principal component analysis outcomes and any patterns with gender.

Results of Research

Science Schools

Correlations

Inter-subject Pearson correlations show very high values, although those for English are slightly lower (Table 1).

Table 1. Pearson correlations (Science Schools).

Subject	Mathematics	Biology	Chemistry	Physics	Social Subjects	Islamic Studies	English
Arabic	0.71	0.78	0.79	0.74	0.68	0.68	0.65
Mathematics		0.75	0.83	0.87	0.71	0.68	0.60
Biology			0.83	0.78	0.73	0.75	0.57
Chemistry				0.85	0.72	0.69	0.62
Physics					0.77	0.73	0.63
Social Subjects						0.78	0.58
Islamic Studies							0.49

Factor Analysis

The scree plot indicates only one factor, accounting for 75% of the variance (a high value). This one factor almost certainly has to be recall and, indeed, looking at the examination papers reveals how the recall of information is all that is required in all subject areas. Table 2 shows the factor loadings, along with means and standard deviations.

Table 2. Factor Loading Analysis (Science Schools).

Science Schools (N = 1333)			
Subject	**Factor Loadings**	**Mean**	**Standard Deviation**
Arabic	0.87	85	9.7
Mathematics	0.90	73	17.7
Biology	0.90	89	10.3
Chemistry	0.92	79	14.5
Physics	0.93	80	12.8
Social Subjects	0.86	92	9.1
Islamic Studies	0.84	94	7.5
English	0.73	78	15.5

Figures 3 and 4 give illustrative histograms (four subjects) showing distributions.

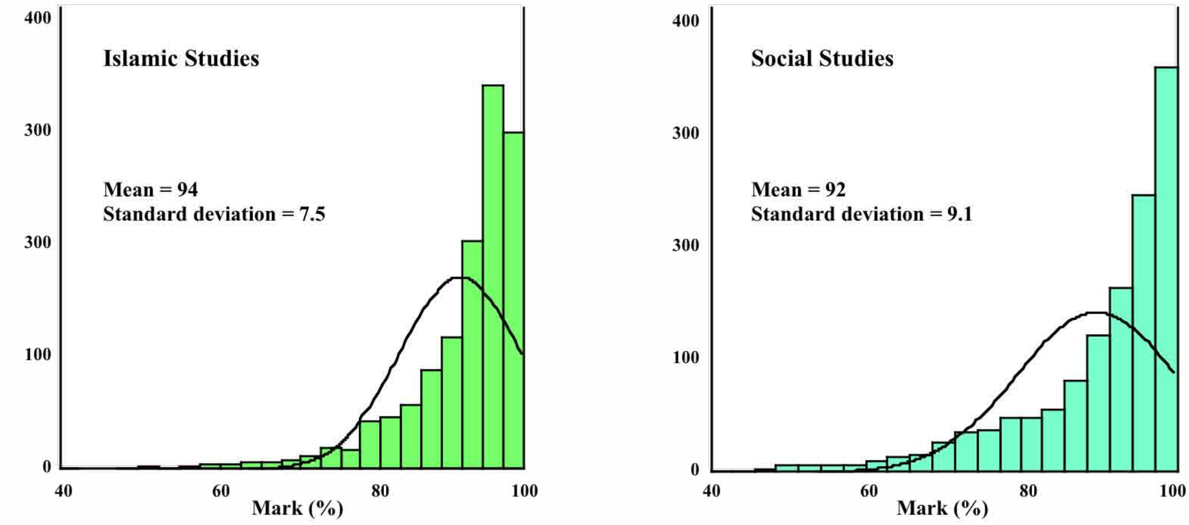

Figure 3: Marks Distributions Islamic Studies and Social Studies (Science Schools).

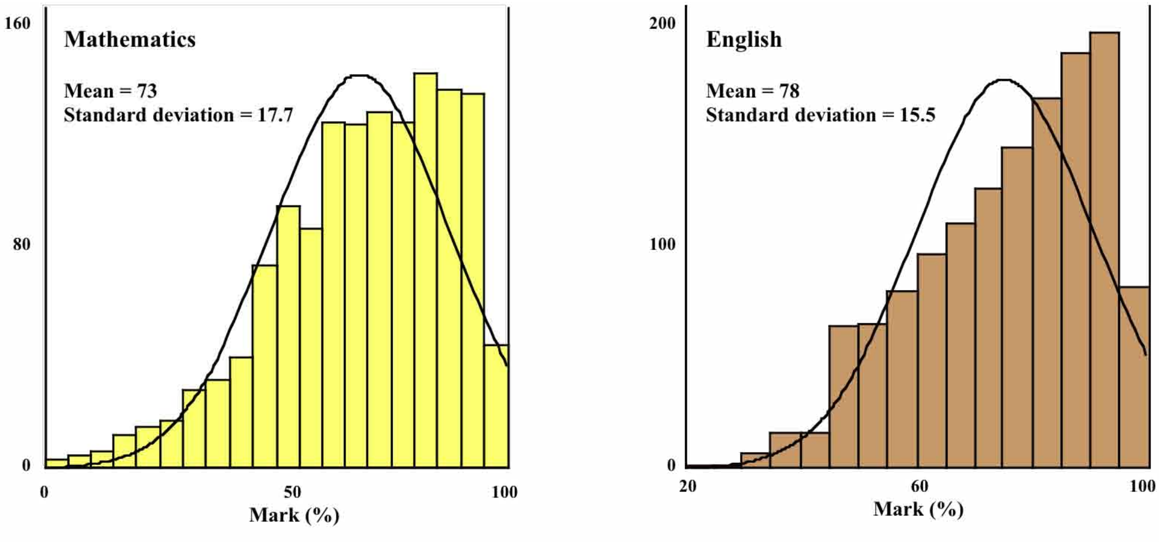

Figure 4: Marks Distributions Mathematics and English (Science Schools).

Gender Comparison

The performance of girls and boys is compared as shown in Table 3.

Table 3. Gender Comparison (Science Schools).

	Girls (N = 1050)		Boys (N = 283)		Comparison	
Subject	Mean	Standard Deviation	Mean	Standard Deviation	t-test	p
Arabic	86.5	9.1	81.1	10.5	7.7	< 0.001
Mathematics	73.1	17.5	70.0	17.8	2.7	< 0.01
Biology	90.1	9.8	86.9	11.7	4	< 0.001
Chemistry	80.6	14.1	75.0	15.0	5.6	< 0.001
Physics	80.1	12.7	80.1	13.3	0.8	n.s.
Social Subjects	92.3	9.2	92.2	8.6	0.6	n.s.
Islamic Studies	93.8	7.3	92.6	8.0	2.3	< 0.05
English	77.4	15.9	78.9	13.8	1.5	n.s.

Where are significant differences, girls always outperform boys (Table 3).

Literary Schools

Correlations

Inter-subject Pearson correlations are shown in Table 4. The correlation values for English are about 0.2 less than those for other inter-relationships.

Table 4. Pearson correlations (Literary Schools).

Subject	Arabic	English	Social Subjects	Contemporary	Environmental
English	0.57		0.55	0.50	0.54
Social Subjects	0.80	0.55		0.79	0.74
Contemporary	0.75	0.50	0.79		0.75
Environmental	0.71	0.54	0.74	0.75	
Islamic Studies	0.73	0.57	0.79	0.78	0.76

Factor Analysis

The Scree plot gives two factors, accounting for 84% of the variance (Table 5).

Table 5. Factor Loading Analysis (Literary Schools).

Literary Schools (N = 803)				
Subject	Factor Loadings		Mean	Standard Deviation
	Factor 1	Factor 2		
Arabic	0.80	0.38	50	9.6
English	0.31	0.95	50	9.3
Social Subjects	0.86	0.31	71	13.4
Contemporary	0.88	0.23	75	13.7
Environmental	0.82	0.31	74	12.3
Islamic Studies	0.87	0.28	51	9.3

The general pattern is very similar to that obtained from the science schools. All the subjects (except English) are simply testing recall.

Gender Comparisons

Table 6. Gender Comparison (Literary Schools).

Subject	Girls (N = 473)		Boys (N = 323)		Comparison	
	Mean	Standard Deviation	Mean	Standard Deviation	t-test	p
Arabic	52.0	9.7	47.0	8.7	7.3	< 0.001
English	49.3	10.0	50.2	8.0	1.3	ns.
Social Subjects	72.0	14.2	68.9	11.9	3.2	< 0.01
Contemporary	74.9	14.1	75.2	13.0	9.3	ns.
Environmental	73.8	13.3	75.3	10.8	1.7	ns.
Islamic Studies	50.7	10.0	51.8	8.1	1.6	ns.

Where there are significant differences, girls outperform boys (Table 6).

Commerce Schools

Correlations

Inter-subject Pearson correlations are shown in Table 7.

Table 7. Pearson correlations (Commerce Schools).

Subject	Arabic	English	Mathematics	Environ-mental	Economics	Banking	Account-ing	Entrepren-eurship	Islamic Education
English	0.63								
Mathematics	0.73	0.62							
Environmental	0.75	0.62	0.76						
Economics	0.75	0.61	0.79	0.80					
Banking	0.74	0.64	0.79	0.78	0.82				
Accounting	0.70	0.61	0.82	0.73	0.76	0.77			
Entrepreneurship	0.60	0.53	0.70	0.72	0.74	0.71	0.70		
Islamic Education	0.69	0.55	0.71	0.75	0.76	0.73	0.70	0.75	
History	0.67	0.60	0.72	0.75	0.75	0.73	0.70	0.78	0.78

In this case, the correlation values are again high.

Factor Analysis

The Scree plot gives only one factor, accounting for 76% of the variance (Table 8).

Table 8. Factor Loading Analysis (Commerce Schools).

Commerce Schools (N = 4886)			
Subject	**Factor Loadings**	**Mean**	**Standard Deviation**
Arabic	0.84	74	11.9
English	0.74	74	14.2
Mathematics	0.89	63	19.9
Environmental	0.89	79	12.7
Economics	0.91	76	14.7
Banking	0.90	82	12.8
Accounting	0.87	78	14.7
Entrepreneurship	0.84	84	12.3
Islamic Education	0.86	84	12.5
History	0.87	82	12.7

The sample here is very large but the pattern of outcomes is very similar to that obtained for the other two groups of schools. Although the means for the examination scores tend to be high, the spread of marks is greater in all subject areas. This can be illustrated by looking at the marks distributions for the two subjects whose means are the highest (Figure 5). Such means make discrimination of ability, especially of the more able, very difficult.

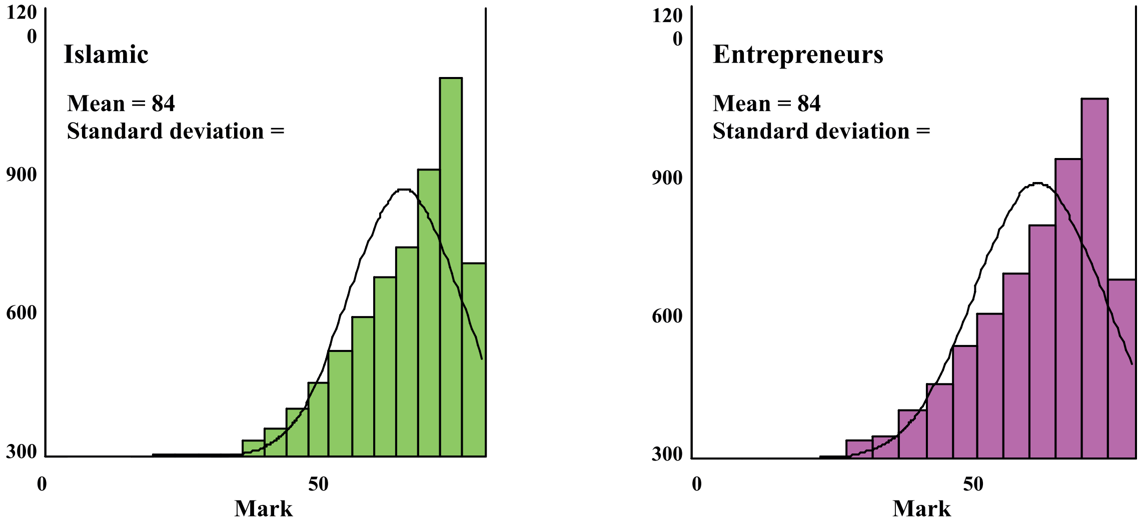

Figure 5: Marks Distributions Islamic Studies and Entrepreneurship.

Gender Comparisons

Table 9. Gender Comparison (Commerce Schools).

Subject	Girls (N = 3049)		Boys (N = 1837)		Comparison	
	Mean	Standard Deviation	Mean	Standard Deviation	t-test	p
Arabic	76.2	11.3	69.8	11.9	18.3	< 0.001
English	74.0	15.0	74.1	12.7	0.1	ns.
Mathematics	65.1	20.3	59.3	18.5	10.0	< 0.001
Environmental	79.3	13.0	77.6	12.0	4.5	< 0.001
Economics	76.7	15.4	75.5	15.5	2.8	< 0.01
Banking	81.8	15.0	81.2	12.4	1.4	ns.
Accounting	80.2	14.0	74.3	15.2	12.8	< 0.001
Entrepreneurship	83.3	12.4	85.6	12.0	6.1	< 0.001
Islamic Education	84.7	12.7	83.7	12.2	2.6	< 0.01
History	81.8	13.4	83.0	11.5	3.4	< 0.001

Girls outperform boys in most subject areas (Table 9).

Discussion

This study arose from a survey of quality in secondary schools in Bahrain (Almadani *et al.,* 2011). In that study, students identified their disquiet about the assessments they had to face. National examinations almost always *control* what teachers do in their own school assessment.

Thus, it could be argued that a key aspect of any quality assurance procedure in any country is to scrutinise the nature of the national assessment, its procedures and its data handling as well as looking at what the examinations are actually assessing. National policies relating to national examinations will control quality in schools simply because teachers and students will focus on the skills where the greatest rewards are available in terms of examination success. Indeed, it could be argued that changing a national examination system would have quite far-reaching effects on every aspect of education in any country. However, this aspect of quality is outwith the control of schools and teachers.

In all three curriculum pathways in Bahrain, the amount of variance '*explained*' by one factor is extremely high. Careful scrutiny of the examinations papers, looking for the ways marks are allocated, reveals that the factor responsible for most of the variance is recall. For example, in the sciences, there is an emphasis on recall of definitions, formulae, facts and taught '*explanations*'. In languages, there is emphasis on recalling grammar, memorized poetry and recall of taught '*interpretations*'. In the business and mathematical areas, much emphasis is given to the recall of facts, information or procedures while, in Islamic Studies and Social Studies, credit is given for recall of memorized text, facts and items of information.

The finding that the factor analyses show that all subjects are measuring the same skill (recall) seems to be a common phenomenon in previous studies in other countries (Al-Ahmadi and Oraif, 2009; Al-Ahmadi and Reid, 2012; Hindal *et al,* 2013). They all showed similar patterns but for much smaller ranges of school subjects. Al-Ahmadi and Reid (2012) were able to show that the factor was *not* related to understanding or the limitations of working memory capacity. They deduced that the one factor was recall and this has been supported by Pidikiti (2007) when he looked at the actual examination papers in a yet another country.

Indeed, in a fascinating study, Bennett (2004) found that the level of problems set in one subject area in the *final* year of university degrees across one country tended to measure the skills only associated with algorithmic problem solving: applying a taught procedure in a routine way. There are numerous far more important skills like understanding, critical thinking, creative thinking, and evaluating that need to play a far greater part in assessment procedures.

Recall may involve recall of information, facts, ideas, or recall of procedures and their correct execution. In the study by Al-Ahmadi and Reid (2012), they employed factor analysis to explore the variables being assessed in a range of measurements. Working with a large sample, they were able to distinguish between recall, understanding, working memory capacity and scientific thinking in their study and different assessments loaded quite precisely on to different factors. Thus, although working memory capacity correlated highly with various examination and test data, working memory capacity loaded precisely on to a *separate* factor. Working memory capacity may *control* performance (Johnstone and Elbanna, 1986, 1989) but it is *not the same thing* as performance.

Rodrigues *et al.* (2010) have revealed some of the ways by which students reach answers. These are often based on factors other than understanding. Indeed, the ways by which students reach answers sometimes do not even depend on any knowledge (and certainly not understanding) of the subject matter at all. Students certainly seem to have devious, albeit legitimate, ways to generate '*correct*' answers by approaches unanticipated by examination setters.

If education is seen as the successful transfer and recall of information, then the assessment system shows success. However, education must involve much more than recall. Indeed, it can be argued that, in an electronic age of '*instant*' information, then recall is relatively unimportant. It is far more important that what is learned is understood. In this, the evidence for understanding involves the ability of being able use knowledge in a novel situation with a good prospect of success. It could be argued that being able to evaluate information is also a vitally important skill. With information available at the press of a button, the ability to evaluate what is useful, valid and reliable is a vital skill.

Knowledge, without the ability to understand and apply, is of limited value. In addition, the development of thinking skills is an important element in all education. Thinking skills have been analysed into critical thinking, creative thinking, scientific thinking and systems thinking (Reid, 2013). When all the knowledge has been forgotten, such skills may remain and prove useful in life in the future.

There is something different with the examinations in English in Bahrain. The inter-subject correlations involving English are lower and, in one of the three curriculum pathways, English loads on to a second factor. Despite looking at the examination papers, it is not obvious what this factor is.

In many subject areas, the actual marks are very high, making discrimination of ability, especially of the more able, very difficult. For example, the average mean in the science curriculum pathway for the 8 subjects is nearly 84% when a mean nearer 60% might be much more appropriate. This reflects a view that marks can be seen to be '*absolute*' (in that a mark of, say 80%, is comparable from year to year or that it indicates a high level of success).

In half of the examinations over all three curriculum pathways, girls outperform boys while boys outperform girls in only two subject areas (entrepreneurship and history). There is a tendency for girls to be more comfortable with memorization, this being observed by Al-Ahmadi (2008) and this may be a possible explanation of the observed differences although, almost certainly, cultural factors will be operating.

Conclusions

This study started from the finding that school secondary students in Bahrain were expressing disquiet about the kinds of assessment they were facing. This is a fundamental aspect of any study on quality assurance: it is essential that the assessment procedures are of quality. Indeed, it has to be recognized that assessment backwashes back on the curriculum, what is taught and how it is taught, all key aspects of quality in education.

This exploration of one national assessment system offers insight on several aspects of assessment that will have powerful effects on the quality of education offered:

- The examinations lack sufficient demand. This is likely to mean that the examinations are unlikely to differentiate well between student ability: the very able are not allowed to '*shine*': thus discrimination is low.
- Correlations between performance across all subjects are extraordinarily high, suggesting a limited range of skills being measured, consistent with some previous studies.
- Factor analytic data show that only ONE skill is, in fact, being measured. A scrutiny of examination papers reveals that is recall, again consistent with some previous studies.
- In the specific setting of Bahrain, examinations in English did not seem to fit exactly the pattern for all other subjects. It would be very interesting to study what it is that English teachers are doing (and examinations are measuring) that is different. There may be important lessons for other subject areas.

There are important implications for examinations systems in all countries. There is a need to measure a much wider range of skills. Indeed, there is a need to specify the nature of agreed desirable skills using descriptions that are operational, enabling assessment to be developed.

However, there are more far-reaching implications in that there is a need to develop assessment techniques and approaches which can be shown to measure some the vitally important educational outcomes which move well beyond recall-recognition.

Assessment is perhaps the single most important factor that will influence quality in education for assessment so often controls the focus and emphasis of the day-to-day teaching and learning at all levels.

Acknowledgements

Appreciation is expressed to the Ministry of Education in the Kingdom of Bahrain which provided the first author with the scholarship and fully paid study leave to pursue this research study in Scotland.

References

ABET, (2003). *International Faculty Workshop for Continuous Program Improvement,* Accreditation Board for Engineering and Technology, Singapore: The National University of Singapore.

Al-Ahmadi, F. (2008). *The development of scientific thinking with senior school physics students,* PhD thesis, University of Glasgow.

Al-Ahmadi, F., Oraif, F. (2009). Working memory capacity, confidence and scientific thinking. *Research in Science & Technological Education, 27* (2), 225-243.

Al-Ahmadi, F., Reid, N. (2012). Scientific Thinking - Can it be Taught? *Journal of Science Education*, *13* (1), 18-23.

Almadani, K., Reid, N., Rodrigues, S. (2011). Quality Assurance: a Pressing Problem for education in the 21st century. *Problems of Education in the 21st Century*, *32*, 9-22.

Angelo, T. (1995). Improving Classroom Assessment to Improve Learning. *Assessment Update*, *7* (6), 13-14.

Assessment Reform Group (2006). *The role of teachers in assessing learning.* Available from CPA Office, Institute of Education, University of London.

Bennett, S. N. (2004). Assessment in Chemistry and the Role of Examinations. *University Chemistry Education*, *8* (2), 52-57.

Broadfoot, P., Black, P. (2001). Redefining Assessment? The first ten years of Assessment in Education. *Assessment in Education*, *11* (1), 7-27.

Hayward, L., Spencer, E. (2006). There is no alternative … to trusting teachers. In Sainsbury, M., Harrison, C. and Watts, A. (Eds) *Assessing Reading - from theories to classrooms,* Maidenhead: NFER.

Hindal, H., Reid, N., Whitehead, R. (2013). A Fresh look at High Ability. *International Journal of Instruction* (paper in press for January 2013).

Johnstone, A. H., El-Banna, H. (1986). Capacity, Demands and Processes - A Predictive Model for Science Education. *Educational in Chemistry, 23*, 80-84.

Johnstone, A. H., El-Banna, H. (1989). Understanding learning difficulties - a predictive research model. *Studies Higher Education*, *14*, 159-68.

Little, A. Ed. (1993). Globalisation, qualifications and livelihoods. *Assessment in Education, 7* (3), 295-312.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd edition), London: Collier Macmillan, 12-103.

National Research Council (2001). *Knowing what Students Know: The Science and Design of Educational Assessment.* National Academy Press, Washington, DC.

Palomba, C. A., Banta, T. W. (1999). *Assessment Essentials: Planning, Implementing, and Improving Assessment in Higher Education.* San Francisco: Jossey-Bass Publishers.

Pidikiti, N. P. (2006). *Performance of secondary students in India related to working memory with reference to some learning styles*, MSc Thesis, University of Glasgow, Glasgow.

Reid, N. (2013). Science Education Research - Ten Key Areas of Findings. *Journal of Science Education,* in press.

Rodrigues, S., Taylor, N., Cameron, M., Syme-Smith, L., Fortuna, C. (2010). Questioning Chemistry: The role of level, familiarity, language and taxonomy. *Science Education International*, *21* (1), 31-46.

Speedy, J., Winter, J., Broadfoot, P., Thomas, J., & Cooper, B. (2003). Researching assessment cultures, researching ourselves. In: R. Sutherland, G. Claxton & A. Pollard (Eds) *Learning and teaching where worldviews meet*. Stoke-on-Trent: Trentham Books, p. 255-271.

Advised by Liviu Moldovan, „Petru Maior" University of Tirgu Mures, Romania

Received: *October 30, 2012* Accepted: *November 29, 2012*

Khaled Ahmed Almadani	PhD Research Student at the School of Education, University of Dundee, Dundee, Scotland, UK. E-mail: K.A.A.Almadani@dundee.ac.uk Website: http://www.dundee.ac.uk/eswce/people/kalmadani.htm
Norman Reid	Professor of Science Education, Universities of Dundee and Glasgow, Scotland, UK. E-mail: dr_n@btinternet.com
Susan Rodrigues	Professor of Science Education, School of Education, University of Northumbria, Coach Lane, Benton, Newcastle upon Tyne, NE7 7XA, England, UK. E-mail: Susan.rodrigues@ed.ac.uk Website: http://www.northumbria.ac.uk/