A paper presented at the 36th Annual Conference of the International Association for Educational Assessment, 22-27 August 2010, Bangkok, Thailand

Evaluating the impact of Bahrain National Examinations: implementing a research agenda

Wafa Al-Yaqoobi, Abdulridha Ali Al-Aradi, Ebby Madera
National Examinations Unit, Quality Assurance Authority for Education and Training, Bahrain
Stuart Shaw
University of Cambridge International Examinations, Cambridge Assessment

E-mail contact:
wafa.alyaqoobi@qaa.edu.bh
abdulridha.alaradi@qaa.edu.bh
ebby.madera@qaa.bh
shaw.s@cie.org.uk

Abstract

Excellence in education is part of Bahrain's vision for 2030. National Examinations were introduced in 2009 to assess the performance levels of students against the national curriculum at key educational stages. The National Examinations Unit (NEU) of the Quality Assurance Authority for Education and Training (QAAET), an independent organisation, was established in 2008 to develop and carry out this work.

With the introduction of any new examinations, awarding bodies are increasingly required to accept responsibility for their impact on major stakeholders. There exists therefore a requirement to make every systematic effort to ensure that the examinations achieve a positive influence on general educational processes and on individuals affected by the examination results.

As part of its commitment to the ongoing review of the quality of its work, the QAAET has proposed (in collaboration with its partner, University of Cambridge International Examinations) a programme of impact research to support claims relating to the overall usefulness of the National Examinations for their intended purpose.

The paper provides an overview of the examinations, a presentation of the theoretical basis of the different impact studies, and a description of the instruments to be used to elicit attitudes, experiences and perceptions of a range of educational beneficiaries.

Key words: Bahrain National Examinations; test impact; high stakes examinations; washback; consequential validity

1. Introduction

With international examinations of ever higher stakes, awarding bodies are increasingly required to accept responsibility for the influence of their tests on a broad range of stakeholders and educational beneficiaries (Weir, 2005; Shaw and Weir, 2007). High stakes tests, so called because they are used to determine admission or otherwise of candidates to specific programmes of study or professions, can be instrumental in shaping educational goals and processes, and society more generally. There exists a requirement, therefore, for any awarding body and other institution to demonstrate and share how it is seeking to meet the demands of validity in its examinations and to make every systematic effort to ensure that its examinations achieve a positive influence or impact on general educational processes and on the individuals who are affected by the results.

As part of the commitment of the Quality Assurance Authority for Education and Training (QAAET) to the ongoing review of the impact and quality of its work, collaborative research

with its international partner the University of Cambridge International Examinations (CIE), is underway in order to investigate the impact of the National Examinations on the educational landscape in Bahrain.

The need to undertake impact studies fits well with growing demands to de-mystify examinations, to make examinations and their development processes as transparent as possible and to respond to the growing acceptance in the educational testing community of the need for ethical approaches to testing (Shohamy, 1997; 1999).

This paper reports on the initial planning, consultation and design phase of a cyclical, iterative impact study of the Bahrain National Examinations. In the first phase of the study, we report on the development of a number of elicitation instruments designed to gather feedback from individuals within one of the main stakeholder groups: the learners and their teachers. The instruments have been designed to determine:

(1) the attitudes, experiences and perceptions of the test taking and teaching population
(2) the wider impact of the National Examinations in Bahrain and their effects on other systems in the administrative and academic contexts of the tests.

It is hoped that the research agenda outlined here will seek to elicit information on the impact of the recently introduced Bahraini National Examinations. A coherent programme of validation and impact research will support claims relating to the overall usefulness of the National Examinations for their intended purpose; explain why the claims are relevant by offering reasons and rationalizations; and supply adequate evidence to support the claims the QAAET wishes to make about its National Examinations.

2. Excellence in education: Bahrain's vision

As part of the Education and Training Reform Project, an initiative of His Royal Highness, the Crown Prince of Bahrain, a decision was taken to ensure that there is quality of education at all levels within the Kingdom of Bahrain. The aim of the project is to help the Ministry of Education, teachers, teacher trainers, and everyone else engaged with learning in schools to raise standards of education in the Kingdom of Bahrain. The Quality Assurance Authority for Education and Training (QAAET) was established by Royal Decree in 2008 and its mandate is to 'review the quality of the performance of education and training institutions in light of the guiding indicators developed by the Authority' (Article 4, 2009). To meet this mandate, four professional units were established within the QAAET: the Schools Review Unit, the Vocational Review Unit, the Higher Education Review Unit, and the National Examinations Unit.

The National Examination Unit (NEU) provides the Ministry of Education and parents with a benchmark for both the performance of the schooling system in Bahrain and the individual performance of students. The NEU provides detailed results and reports to the Ministry of Education. In these reports the results are broken down by student, by class, by school and by year. The reports also provide a breakdown of results by topics and skills from the Ministry's subject curricula.

During the NEU's first year of operation, national examinations for Grades 3 and 6 were conducted in all government Primary and Primary-Intermediate schools in May 2009 for the first time. The first year of examinations helped serve to establish a baseline against which future performances can be measured. A total of approximately 21,000 students took the examinations, which in Grade 3 were Arabic and Mathematics, and in Grade 6 were Arabic, Mathematics, Science and English. In all subjects the examinations covered the whole curriculum. Because of the recognition of the importance of these subjects by other national exams around the world, choosing them allows benchmarking with regional or international

standards All examinations were marked in Bahrain by teachers working in Bahraini government schools, and results were published to schools and students in June 2009.

Because they are taken every year, the National Examinations will provide a way of checking improvements in schools and in the educational system as a whole over a period of time. Students and teachers will have reliable information about the strengths of their performance in different subjects, and this will help them to see where their strengths are – as well as where there is most room for improvement.

3. Primary purposes of the National Examinations

The main purpose of the National Examinations is to provide information about the performance of students on broad divisions of subject content based on the competencies in the National Curriculum and to provide information on performance on different question types. The information will contribute to the work of the Ministry of Education in formulating policy; the work of the Directorate of Curricula in developing the National Curriculum and its associated textbooks; the work of the Centre for Measurement and Evaluation in evaluating textbooks and teaching strategies; the work of the Teacher Training College in targeting initial and in-service teacher training; and, the work of the School Review Unit of the QAA in its reviews of schools.

Other purposes of the National Examinations include the provision of robust information for the monitoring of standards over time and for research into value added.

An objective of the NEU is to facilitate the spread of good practice throughout the Bahraini education system. In light of this, it is hoped that the National Examinations will provide a beacon of good assessment practice.

3.1 Reporting student performance

Student performance on the National Examinations is reported in four ways: student reports, teaching group reports, school reports, and Ministry of Education reports.

The results are given in two forms: a *normalized percentage score* and a *performance score*, both of which indicate the standard of performance in a subject.

Normalized percentage score
The normalized percentage score, calculated by defining the national average score as 70% and the standard deviation as 10%, is designed:

- to show how a student's performance compares with that of the other students in the same year and in the same grade;
- to compare students' National Examination percentage score with their percentage marks in internal school run tests.

Performance score
The performance score is calculated from students' abilities on a Rasch scale (from 0.0 to 8.0). The national average was defined as 4.0 in the first year of testing and subsequent years' tests will be anchored to the scale that was set in the first year. (The national average is not always the same.) The performance score is useful:

- for tracking the performance of students between grades 3 and 6, or between grades 6 and 9;
- for monitoring trends in performance from year to year;
- for comparing students' performance in different areas and different skills within a subject.

4. Approach to the study of test impact

Within the broader educational measurement literature, the impact concept includes the wider influences of tests on the stakeholder community (Hamp-Lyons, 1997; Wall, 1997 ; Weiss, 1998; McNamara, 2000; Saville, 2009), in other words, the influences of educational assessments – their effects and consequences, beyond the immediate learning context (Hawkey, 2006). Thus impact research investigates issues of test use and social impact in *macro* contexts of society (Bachman and Palmer, 1996; Saville, 2009). Impact is generally considered to include 'washback', the effect of testing on teaching and instruction (Davies et al., 1999, p.225) and is seen as "one form of impact" (Hamp-Lyons, 1997, p.299). If impact relates to test use at the macro level, then the more traditional concept of washback relates to narrower *micro* contexts in the learning environment (the classroom, school, micro level participants).

The dynamic relationship between the impact and washback contexts reflects the growing importance of the need for awarding bodies to implement an extensive test validation, evidence-based approach to educational planning and to the development of high-stakes tests particularly where the tests have widespread currency and recognition. "Washback and the impact of tests more generally has become", therefore "a major area of study within educational research" (Alderson, 2004, p.ix) and related studies can play a significant role in ensuring that ethical testing is achieved. For example, Shohamy (1999) notes that test instruments "are powerful because they lead to momentous decisions affecting individuals and programs" (1999, p.4). Hamp-Lyons argues that an awarding body should evaluate its test instruments "from the perspective not only of the test-setter but also of the other stakeholders" (1997, p.299), the stakeholders being all those who have a legitimate interest (direct or indirect) in the use or effect of a particular assessment or its evaluation (Weiss, 1998, p.337; see also Rea-Dickins,1997, p.305, for a comprehensive list of potential stakeholders in educational testing).

The impact concept used here is grounded in a highly progressive, 'consequentialist' conception of validity. The concept covers in part what Messick (1989) terms *consequential validity*, arguing that it is necessary in validation studies to ascertain whether the social consequences of test use and interpretation support the intended testing purpose(s) and are consistent with other social values. The extent to which the inferences which are made on the basis of the outcomes of an examination are meaningful, useful and appropriate is regarded as an essential aspect of validity (Cambridge Assessment, 2009, p, 8). Score-based inferences are perceived by Messick as a function of the external social consequences of the testing who stresses that " … the social values and the social consequences [of an examination] cannot be ignored in considerations of validity" (Messick 1989, p.20).

Shohamy (1993, p.37) argues that "testers must begin to examine the consequences of the tests they develop… often... they do not find it necessary to observe the actual use of the test." Similarly, Messick (1996, p.247) questions whether… "the scores have utility for the proposed purposes in the applied settings. Are the short and long term consequences of score interpretation and use supportive of the general testing aims and are there any adverse side effects?"

The consequentialist view enables awarding bodies not only to focus on whether an assessment is measuring what it is intended to measure, but also to provide a focus on whether the outcomes of the assessment are being used in an appropriate way. Whilst an awarding body cannot be held responsible for all possible uses of the outcomes of the assessments it provides, it can take responsibility for being very clear regarding legitimate uses, can investigate patterns of use and impact, and issue cogent and precise guidance to stakeholders on how to interpret test scores. A consequentialist approach also helps define what should be incorporated in validation studies.

The study of test impact is one of several elements in a continuous and iterative test validation process (Shohamy, 1999:20). Collecting clear, explicit, and comprehensively specified validity evidence in support of the claims about an examination's suitability for its intended purpose is, therefore, a paramount concern for any validation exercise (Weir, 2005, p.47)

It is argued here that impact studies that periodically monitor context of learning and context of test use should be undertaken in a cyclical, iterative manner such that they inform future test development and policy making.

The importance ascribed by Cambridge Assessment, for example, to impact studies is well documented both in the context of language testing (Green 2003; Weir and Milanovic 2003; Hawkey 2006; Saville 2009); and also in the area of general education (Beedle, Eason, and Maughan, 2007; Shaw and Allen, 2009).

5. The National Examinations impact study

The QAAET has identified the need to monitor the effects of their National Examinations on a diverse range of stakeholders by eliciting their views, perspectives and attitudes. They have also recognised the requirement to explore the effects of the examinations on teaching materials and classroom activity.

Saville (2003) describes a set of procedures for collecting evidence that allows impact to be estimated. Each procedure represents a potential point in the impact research programme for data collection:

- who is taking the examination;
- who is using the examination results and for what purpose;
- who is teaching towards the examinations and under what circumstances;
- what kinds of courses and materials are being designed and used to prepare candidates;
- what effect the examinations have on public perceptions generally (e.g. regarding educational standards);
- how the examinations are viewed by those directly involved in educational processes (e.g. by students, examination takers, teachers, parents, etc.); and,
- how the examinations are perceived by members of society outside education (e.g. by politicians, businessmen, Ministries of Education, etc.).

Clearly, the extent and scale of the research programme will depend on the availability of data, the practicality (or impracticality) of collecting the data, the engagement and interests of stakeholders, and any specific research questions (which themselves will be dictated by the context, purpose, aims and objectives of the impact studies). As Hawkey (2006) points out, the issues need to be considered carefully in terms of the costs of doing impact research and any benefits emerging from that research.

In the context of the Bahrain National Examinations, specific research questions include:

- What are experiences/perceptions of students who have taken the National Examinations?
- What is the impact of the National Examinations on students who are preparing for them or have taken them?
- Who is teaching towards the National Examinations and under what circumstances?
- What is the 'washback' of the National Examinations on courses preparing students?
- What kinds of courses and materials are being designed and used to prepare students for the National Examinations?
- Who is using the results from the National Examinations and for what purpose?

Investigating impact is regarded as being an essential aspect of determining the utility (or usefulness) of an educational assessment in terms of fulfilling its intended purpose, that is, its fitness for specific purposes and contexts of use. As impact studies play an integral role in assessment accountability (more particularly through the validation process), the logical starting place for impact evaluation is a clear statement of the proposed interpretations and purposes of the Bahrain National Examinations (Kane, 2006, p.23).

According to Kane (2006), evidence required for validation depends on the claims being made. An integral feature of the validation process must entail, therefore, collecting adequate evidence to support the examination board's claims about the examination's suitability for its intended purpose. However, providing appropriate evidence for validity is not a simple undertaking.

It is important to the success of the Bahraini impact research that a suite of standardised data collection instruments and procedures (such as questionnaires and interview schedules) are developed in order to understand the test impact better and to conduct effective surveys to monitor it (Hawkey, 2004). The instruments for this study are adaptations of existing impact instruments and include:

- **modular student questionnaires** including questions relating to student perceptions, views and attitudes; affective and motivational aspects.
- **modular teacher questionnaires** including questions relating to demographic and biographical detail; perceptions of examinations and student performance; nature and content of courses provided; extent and use of materials; curricular issues.
- **school lesson observation analysis instruments** enabling data to be collected on lesson types; materials; facilities; student activities; communicative opportunity analysis information.
- **semi-structured interview/focused discussion group protocols** for students and teachers covering aspects relating to background; learning/study approaches and teaching approaches; perceptions; teaching and assessment issues; test validity and reliability.
- **receiving institution questionnaires** for completion by Higher Education admissions and teaching staff eliciting their views on how examination results are used; predictive validity and reliability; impact; tertiary level preparation/success.
- **textbook and course materials instruments** enabling data to be collected on teacher/student perceptions and enabling evaluations of existing preparatory materials and curriculum.

The design of the impact research agenda shown here illustrates the cyclical, iterative nature of impact studies:

Phase 1: Initial planning, consultation and design (August 2009 to July 2010)

- conceptualisation
- stakeholder consultations with NEU/QAAET
- development of project plan and validation strategy
- adaptation, development and pre-piloting of qualitative instruments (student and teacher questionnaire)
- small-scale data collection and analysis
- determine ambiguous, invalid, unreliable questionnaire items
- refine instruments for piloting on wider scale
- dissemination of preliminary issues (IAEA 2010)

Phase 2: Development (August 2010 to July 2011)

- refine research hypotheses and questions
- refine student and teacher questionnaires
- establish pilot sampling criteria
- pilot student and teacher instruments on wider scale
- data collection and analysis
- develop lesson observation instruments, participant protocols, Higher Education elicitation and washback instruments
- dissemination of findings (conference and publications)

Phase 3: Implementation and validation (August 2011 to July 2012)

- pilot instruments developed in Phase 2
- larger-scale data collection/analysis
- increased stakeholder engagement (consulting and informing)
- validation studies – triangulation of data
- dissemination of findings (conference and publications)

Phase 4: Extended data collection (August 2012 onwards)

- full-scale implementation of all data collection instruments
- further refine methodology for studying test impact in societal contexts
- informing stakeholders of study findings
- dissemination of findings (conference and publications)
- raise profile of research undertaken by NEU

6. Lessons learnt from pre-pilot findings

This paper reports on Phase 1 of the study, in particular, the issues emerging from pre-piloting the student and teacher questionnaires (see Baker, 1994) for a discussion of issues relating to pre-testing/piloting of a particular research instrument.

The questionnaires were initially constructed through a series of meetings and consultations with expert participants with relevant research and background experience. The instruments were then reviewed for appropriateness and adapted for the Bahrain context before being translated into Arabic.

A formal request was sent from the QAAET to the Ministry of Education requesting permission to administer the questionnaire to students in schools. Pre-piloting was carried out in April, 2010. Students who had taken the National Examinations in 2009 constituted the target population for the pre-pilot. The first National Examinations were administered in 2009 to grade 3 and grade 6 students (ages 7-8 and 10-11 respectively). Therefore, at the time of the pre-pilot, the target population included those students who were in grades 4 and 7.

Piloting provides a “clear definition of the focus of the study” (Frankland and Bloor, 1999, p. 154) and enables the collection of data to be focused on a restricted range of analytical topics. Pre-piloting the questionnaire helps determine whether the questions as they are worded will achieve the desired results; whether the questions will be understood by all classes of respondent (this proved to be an important issue for Grade 3 students), and whether additional or specifying questions would be needed or whether some questions should be eliminated (De Vaus, 1993).

The respondents (teachers and students) selected for the pre-pilot survey were broadly representative of the type of respondent to be interviewed in the main pilot and live surveys. The pre-pilot sample comprised four schools (two boys’ schools and two girls’ schools). The

distribution of students and teachers by Governorate, Gender and Age for both grades is given in Tables 1 and 2 respectively.

Table 1: Distribution of students by Governorate, Gender and Age for Grades 3 and 6

Governorate	Gender	Grade 3 (age 8-9)	Grade 6 (age <11)	(age 11-12)	(age 13-14)	Total
Capital	male	14				14
Northern	female	15				15
Muharraq	male			14	2	16
	female		3	12	0	15
					Total	60

Table 2: Distribution teachers by Governorate, Gender and Age for Grades 3 and 6

Governorate	School	Teacher Gender	Grade 3	Grade 6	Total
Capital	boys	female	5		5
Northern	girls	female	5		5
Muharraq	boys	male		5	5
	girls	female		5	5
				Total	20

Pre-pilot exploratory data was collected by four education specialists working within the NEU ('interviewers'). Each interviewer visited one school. Prior to administering the questionnaires, the interviewers sought the consent of the students by allowing them to read and sign a consent form.

In general, pilot studies can be "time-consuming, frustrating, and fraught with unanticipated problems" (Mason and Zuercher, 1995). Moreover, pilot studies are often "underdiscussed, underused and underreported" (Prescott and Soeken, 1989, p.60). A brief review of the research literature, for example, would suggest that pilot studies are seldom reported in any detail and rarely in full (van Teijlingen et al. 2001). Despite these difficulties a pilot study can reveal deficiencies in the design of a proposed research procedure indicating where the main research could be jeopardised and whether the proposed approaches or instruments are too complicated or simply inappropriate. It is important, therefore, to report issues which emerge from pre-piloting activities particularly if they affect the direction and nature of the research programme.

Quite apart from the limitations normally associated with pilot studies (such as the possibility of making inaccurate predictions or assumptions on the basis of pilot data; difficulties arising from contamination; etc.), a number of potentially challenging yet interesting Bahrain-specific issues were identified during pre-pilot data collection. These issues offer advanced warnings into:

- the efficacy and feasibility of future full-scale surveys;
- the requirement for a clear articulation of, and adherence to, a research protocol (and whether the research protocol is in fact realistic and workable);
- whether the proposed instruments are too complex or inappropriate;
- identification of logistical problems which may arise during the implementation of the proposed instruments;
- identification of potential practical problems in following the research procedure;
- determining what resources will be needed for a full-scale live survey;
- managing local policy and practice (such as seeking permission to gain access to participants)

Decisions relating to refining research hypotheses and questions together with subsequent improvements to instruments (Phase 2 of the programme) must take account of a number of practical concerns.

Firstly, it is not uncommon for teachers in schools in Bahrain to read questions to students. This is especially true for younger students (Grades 1-3). NEU interviewers also found it necessary to read, and in some cases explain, questions to pre-pilot students. This raises questions about whether the process of responding to the questions is fundamentally altered by the requirement to describe and explain the questions to respondents. 'Interviewer bias' constitutes a serious threat to reliability (a characteristic of the testing instrument) and also validity (the way the testing instrument is employed). For example, a 'response effect' can arise as a result of the eagerness of the students to please the interviewer (especially pronounced among younger students) or from a tendency exhibited by the interviewer to orchestrate those answers that support preconceived notions.

The interviewers freely interacted with students when discussing questions and encouraged them to 'write something' on the questionnaire. As a consequence, it was occasionally necessary when evaluating questionnaire responses to discard certain anecdotal evidence thought to be considered misleading.

The practice of reading questions was both unanticipated and compounded by the fact that each of the four interviewers applied different protocols during administration of the questionnaire. Clearly, feedback from students to identify ambiguities and difficult questions needs to be captured in a systematic and uniform manner. For future surveys it will be necessary to standardize the questionnaire administration process by using a set of agreed protocols.

Secondly, the majority of students across both grades did not understand the questions. Grade 4 students were particularly unfamiliar with some of the words used on the questionnaire and sought clarification from teachers and NEU interviewers. Ambiguities might be the result of linguistically demanding questions, insufficient first language proficiency, or inadequately translated questions. There is a clear requirement to investigate and establish the levels of language at which the participants are functioning and construct questionnaires that match these levels of linguistic competence. This will facilitate student access to the questionnaires and remove any interviewer influences.

A third issue relates to the administration of a translated questionnaire through an 'interlocutor'. Every effort was made during translation to make the Arabic version equivalent to its English counterpart. However, there was no independent verification of the process through back translation to English to establish equivalence. This was further exacerbated by the interviewers who acted as interlocutors during the pre-pilot. One way of addressing this issue is to undertake an independent back translation to English and verify the direct equivalence of the translated version.

Another issue relates to the timing of questionnaire administration. The time between students actually taking the National Examinations and administration of the questionnaire proved to be too long. Many students were unable to remember their experience of taking the examinations and struggled to make meaningful contextual connections between the questionnaire questions and their own recalled perceptions.

If the student questionnaire in the main study is to be administered in exactly the same way as in the pilot then this will have significant logistical implications, for example, access to participants, time, personnel resources, and disruption to educational programmes and student activities.

7. Conclusion: the way forward

This paper has reported on work undertaken during the first phase of the Bahrain impact study programme (*Initial planning, consultation and design*). Impact studies of the type described here and which investigate the positive influence on general educational processes and on educational beneficiaries directly affected by the test results perform a hugely important role in test accountability.

It is hoped that the issues identified in the first stage of the study more generally (such as insufficient or problematic piloting; interviewer effects), and the pre-piloting activity in particular, will add to the already growing body of lessons learnt from previous educational impact research. Hawkey (2006) provides a helpful overview of the real and anticipated difficulties in impact research which include:

- a lack of clear and stated aims
- implicit rather than explicit theoretical input
- inability to establish causal relationships
- inadequate sampling as a result of difficulty in gaining access to participants
- insufficient piloting, thus instruments containing ambiguous, invalid, unreliable items
- lack of triangulation through other data collection methods
- interviewer or researcher effects

Further work is necessary to modify the existing survey questionnaires in light of the pre-pilot before a more extensive larger scale study in Phase 2 can be undertaken. It is evident from the findings so far that the second *Development* phase of the impact research programme will require a significant investment of resources especially in relation to the generation of other impact elicitation instruments. Given the findings from the pre-pilot it may be necessary to adopt a conceptually different approach to the re-construction of existing impact instruments and to the construction of new ones in the context of Bahrain.

Areas of future research should focus on:

- refining of the methodology for studying test impact in societal contexts;
- ensuring reliability and validity of the impact instruments (giving consideration to aspects of content, predictive, or construct validity); and
- pilot and main sample considerations including framing, sample size, power calculations, and contamination issues.

8. References

Alderson, J. (2004). Introduction, in Cheng, L and Watanabe, Y (eds.) *Context and Method in Washback Research: The influence of language testing on teaching and learning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bachman, L. and A. Palmer. (1996). *Language Testing in Practice*. Oxford: OUP

Beedle, P., Eason, T. and Maughan, S. (2007). *A Case Study of the Development of an International Curriculum Leading to International GCSE Certication.* In The SAGE Handbook of Research in International Education. (Eds. Hayden, M., Levy, J and Thompson, J). London: SAGE Publications.

Cambridge Assessment. (2009). The Cambridge Approach: principles for designing, administering and evaluating assessment. Cambridge: A Cambridge Assessment Publication.

Davies, A, Brown, A, Elder, C, Hill, K, Lumley, T and McNamara, T. (1999). *Dictionary of Language Testing, Studies in Language Testing 7*. Cambridge: UCLES and Cambridge University Press.

De Vaus, D.A. (1993). *Surveys in Social Research* (3rd edn.), London: UCL Press.

Frankland, J. and Bloor, M. (1999), Some issues arising in the systematic analysis of focus group material, In: Barbour, R. and Kitzinger, J. (eds) *Developing Focus Group Research: Politics, Theory & Practice*, London: Sage

Green, A. (2003). *Test Impact and EAP: a comparative study in backwash between IELTS preparation and university pre-sessional courses.* Research for the Ph.D degree, University of Surrey at Roehampton.

Hamp-Lyons, L. (1997). Washback, impact and validity: ethical concerns. *Language Testing, 14*, 295-303.

Hawkey, R. (2006). The theory and practice of impact studies: Messages from studies of the IELTS test and Progetto Lingue 2000, *Studies in Language Testing*, Vol 24, Cambridge: Cambridge ESOL and Cambridge University Press.

Hawkey, R. (2004). An IELTS Impact Study: implementation and some early findings. *Research Notes 15*.

Kane, M.T. (2006). Validation. In *Education measurement* (4th ed.), ed. R.L. Brennan. Westport: Praeger.

McNamara, T. (2000). *Language Testing* Oxford:OUP

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13, 3, 241-256.

Messick, S. (1989). Validity. In R. Linn (Ed.) *Educational measurement* (pp. 13-103). New York: Macmillan.

Prescott, P.A. and Soeken, K.L. (1989), The potential uses of pilot work. *Nursing Research* 38: 60-62.

Rea-Dickins, P. (1997). So why do we need relationships with stakeholders in language testing? A view from the UK, *Language Testing 14* (3), 303-314.

Saville, N. (2009) Language assessment in the management of international migration: A framework for considering the issues, *Language Assessment Quarterly*, 6(1): 17-29.

Saville, N. (2003). The process of test development and revision within UCLES EFL. In Milanovic and Weir, eds. *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913-2002*. Cambridge: CUP/UCLES.

Shaw, S. D. and Allen, L. (2009). *IGCSE Impact Study: the test takers' perspectives.* Cambridge International examinations internal report.

Shaw, S.D., and C.J. Weir. 2007. *Examining writing: research and practice in assessing second language writing*. Vol. 26 of *Studies in Language Testing*. Cambridge: UCLES/CUP.

Shohamy, E. (1999). *Language Testing: Impact*. In Spolsky, B (ed.) Concise Encyclopedia of Educational Linguistics, Oxford: Pergammon

Shohamy, E. (1997). Testing Methods, Testing Consequences: Are they ethical? Are they fair? *Language Testing* 14, 3: 340-349.

Shohamy, E. (1993). *The Power of Tests. The Impact of Language Tests on Teaching and Learning.* Washington, DC: NFLC Occasional Papers

Teijlingen van, E., Rennie, A.M., Hundley, V., Graham, W. (2001), The importance of conducting and reporting pilot studies: the example of the Scottish Births Survey, *Journal of Advanced Nursing* 34: 289-295.

Wall, D. (1997). Test Impact and Washback, in Language Testing and Evaluation, Volume 7 of Kluwer *Encyclopedia of Language Education*, 291–302.

Weir, C.J. (2005). *Language testing and validity evidence*. London: Palgrave.

Weir, C J and Milanovic, M (Eds) (2003) *Continuity and Innovation: The History of the CPE 1913-2002*. Studies in Language Testing 15, Cambridge: Cambridge University Press.

Weiss, C. (1998). *Evaluation*. New Jersey: Prentice Hall.