Language Assessment for Multilingualism

Proceedings of the ALTE Paris Conference, April 2014

For a complete list of titles please visit: www.cambridge.org/elt/silt

Also in this series:

European Language Testing in a Global Context: Proceedings of the ALTE Barcelona Conference July 2001
Edited by Cyril J. Weir and Michael Milanovic

IELTS Collected Papers: Research in speaking and writing assessment
Edited by Lynda Taylor and Peter Falvey

Changing Language Teaching through Language Testing: A washback study
Liying Cheng

The Impact of High-stakes Examinations on Classroom Teaching: A case study using insights from testing and innovation theory
Dianne Wall

Impact Theory and Practice: Studies of the IELTS test and *Progetto Lingue 2000*
Roger Hawkey

IELTS Washback in Context: Preparation for academic writing in higher education
Anthony Green

Examining Writing: Research and practice in assessing second language writing
Stuart D. Shaw and Cyril J. Weir

Multilingualism and Assessment: Achieving transparency, assuring quality, sustaining diversity – Proceedings of the ALTE Berlin Conference, May 2005
Edited by Lynda Taylor and Cyril J. Weir

Examining FCE and CAE: Key issues and recurring themes in developing the First Certificate in English and Certificate in Advanced English exams
Roger Hawkey

Language Testing Matters: Investigating the wider social and educational impact of assessment – Proceedings of the ALTE Cambridge Conference, April 2008
Edited by Lynda Taylor and Cyril J. Weir

Components of L2 Reading: Linguistic and processing factors in the reading test performances of Japanese EFL learners
Toshihiko Shiotsu

Aligning Tests with the CEFR: Reflections on using the Council of Europe's draft Manual
Edited by Waldemar Martyniuk

Examining Reading: Research and practice in assessing second language reading
Hanan Khalifa and Cyril J. Weir

Examining Speaking: Research and practice in assessing second language speaking
Edited by Lynda Taylor

IELTS Collected Papers 2: Research in reading and listening assessment
Edited by Lynda Taylor and Cyril J. Weir

Examining Listening: Research and practice in assessing second language listening
Edited by Ardeshir Geranpayeh and Lynda Taylor

Exploring Language Frameworks: Proceedings of the ALTE Kraków Conference, July 2011
Edited by Evelina D. Galaczi and Cyril J. Weir

Measured Constructs: A history of Cambridge English language examinations 1913–2012
Cyril J. Weir, Ivana Vidaković, Evelina D. Galaczi

Cambridge English Exams – The First Hundred Years: A history of English language assessment from the University of Cambridge 1913–2013
Roger Hawkey and Michael Milanovic

Testing Reading Through Summary: Investigating summary completion tasks for assessing reading comprehension ability
Lynda Taylor

Multilingual Frameworks: The construction and use of multilingual proficiency frameworks
Neil Jones

Validating Second Language Reading Examinations: Establishing the validity of the GEPT through alignment with the Common European Framework of Reference
Rachel Yi-fen Wu

Assessing Language Teachers' Professional Skills and Knowledge
Edited by Rosemary Wilson and Monica Poulter

Second Language Assessment and Mixed Methods Research
Edited by Aleidine J Moeller, John W Creswell and Nick Saville

Language Assessment for Multilingualism

Proceedings of the ALTE Paris Conference, April 2014

Edited by

Coreen Docherty
Cambridge English Language Assessment

and

Fiona Barker
Cambridge English Language Assessment

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781316505007

First published 2016

Printed in XXXX

A catalogue record for this publication is available from the British Library

Library of Congress Cataloging-in-Publication Data
Names: ALTE Conference (5th : 2014 : Paris, France) | Docherty, Coreen, editor. | Barker, Fiona, editor.
Title: Language assessment for multilingualism : Proceedings of the ALTE Paris Conference, April 2014 / edited by Coreen Docherty, Cambridge English Language Assessment and Fiona Barker, Cambridge English Language Assessment.
Description: Cambridge ; New York : Cambridge University Press, [2016] | Series: Studies in Language Testing ; 44 | Includes bibliographical references.
Identifiers: LCCN 2015034884 | ISBN 9781316505007
Subjects: LCSH: Language and languages--Ability testing--Europe--Congresses. | Second language acquisition--Ability testing--Europe--Congresses.
Classification: LCC P118.75 A482 2014 | DDC 404/.2--dc23 LC record available at http://lccn.loc.gov/2015034884

15 Teacher involvement in high-stakes testing

Daniel Xerri
Patricia Vella Briffa
University of Malta

Abstract

This paper explores the premise that teachers' involvement in high-stakes testing is desirable because the resulting test is a product of their knowledge of the learning context, the student cohort, and the subject content. Such involvement is indicative of an increased sense of trust in teachers' judgements. By means of a case study approach, this paper discusses the process of developing a public examination from the authors' combined perspectives as researchers and teachers whose assessment literacy was enhanced because they were privileged to be involved at every stage. This paper outlines the challenges faced and elaborates on the lessons learned from their prolonged involvement. It evaluates the implications of teachers' involvement in high-stakes testing and seeks to contribute to a better understanding of the benefits that may arise when teachers are invited to play an instrumental role in the design and implementation of such examinations.

Introduction

Teachers' involvement in high-stakes test development can enhance their assessment literacy and result in examinations that are informed by their knowledge of the learning context, the student cohort, and the subject content. There is a dearth of research on how teachers' involvement in public examinations translates into such potential benefits. The idea that teachers should be encouraged to don the examiner's hat has not been given sufficient attention in the assessment literature. In fact, Sasanguie, Elen, Clarebout, Van den Noortgate, Vandenabeele and De Fraine (2011:908) point out that 'Despite [high-stakes tests'] great impact, discussions on the separation versus combination of teaching and assessment roles are rare and empirical research is nearly absent'. This paper therefore sheds light on the benefits that may be derived when teachers actively contribute to high-stakes examinations.

In this paper we present a case study of our involvement in high-stakes

testing by evaluating our role as teachers in the design and implementation of a newly introduced English speaking component forming part of a popular public examination at Advanced level in Malta used for university admissions. Our experience as teachers allowed us to identify and address the gaps present in the syllabus in order for our students and other candidates to be provided with a reliable and valid form of assessment of their speaking skills at this advanced stage of language learning. This paper analyses our contribution to this speaking examination from its inception up to the first sitting by a national cohort of candidates. By demonstrating what we learned from a three-year process made up of a number of test development stages, this paper illustrates how teachers' involvement in public examinations could help develop their assessment literacy and lead to a more equitable form of high-stakes testing.

Concerns with high-stakes testing

The impact of language tests can be far-reaching, especially if these tests are of a high-stakes nature. Taylor (2005:2) affirms that 'the use of tests and test scores can impact significantly on the career or life chances of individual test takers'. Over the past few years a number of countries seem to have placed a stronger emphasis on high-stakes testing. A case in point is the USA where high-stakes testing is becoming the chief means of assessing students and gauging teacher and school accountability. However, high-stakes testing receives a fair amount of criticism, especially because it is accused of reproducing social and educational inequality (Au 2008) and for being mechanistic and reductive (Allen 2012). It does so by binding academic success to performance on tests that might be based on a limited set of measurable outcomes to the exclusion of other significant areas of learning. Grant (2004:6) labels high-stakes tests 'oppressive' because they impair quality teaching and learning, subject students to a restricted curriculum, and push teachers to teach to the test. In their research on the impact of a school-leaving English examination in Poland, Lewkowicz and Zawadowska-Kittel (2008:30) found that teachers focus on task types that feature in examinations and teach students strategies that enable them to do well on a test. Similarly, a study focusing on the Nigerian context found that a preoccupation with attainting certification has promoted teaching and learning oriented primarily towards passing the test rather than enhancing language use (Christopher 2009:12). Nichols's (2007:57) review of the literature on the impact on student achievement of high-stakes tests leads her to posit that 'the findings from the most rigorous studies on high-stakes testing do not provide convincing evidence that high-stakes testing has the intended effect of increasing student learning'. In fact, the unintended outcomes of high-stakes testing are largely negative, especially on instruction and on teacher and student motivation (Jones 2007).

However, high-stakes tests have become an intrinsic part of the contemporary educational milieu and they can have a positive washback effect on teaching and learning. Hence, it might be better for teachers to use them to their advantage rather than seeking to debunk them at every turn. In our case, we argued that it would be more profitable for us to be involved in a high-stakes examination rather than distancing ourselves from it and complaining about its effects.

High-stakes testing can affect teachers in a number of ways, especially if they are made to feel that they have no sense of ownership over the test or that it is exclusively determining the nature of teaching and learning. Currently, the driving force behind the curriculum that teachers focus on in class seems to be constituted by 'the pressures of assessment systems that pay little heed to consistency or coherence between teachers' visions of desirable education and those articulated in high-stakes examinations' (Atkin 2007:57). These pressures can impinge on classroom practice, stifle teachers' views and make them feel disenfranchised (Nichols and Berliner 2007). This is especially so when teachers are not given the opportunity to be involved in the development of high-stakes tests. High-stakes testing can lead teachers to 'increasingly feel that they are at the mercy of forces beyond their control' (Reich and Bally 2010:181). For example, Costigan (2002:32) reports that the amount of high-stakes testing that a small group of primary school teachers were faced with when they entered the profession not only affected the quality and type of instruction they delivered but also made them feel disempowered. Focusing on public school teachers in New York City, Crocco and Costigan (2006:1) contend that 'high-stakes testing has produced high-stakes teaching in many schools, raising the risk of aggravating the already high level of teacher attrition'. Such assessment-driven teaching burdens teachers with undue pressure. A study by Assaf (2008:249) shows how an English Language Learner (ELL) reading teacher struggled to act autonomously due to testing pressures and felt forced to reinvent her professional identity so as to be in synch with the testing culture in her context. This is in line with studies indicating that the pressure of high-stakes testing might lead teachers to change their instructional practices (Hoffman, Assaf and Paris 2001) and affects the way they respond to students' learning needs (Flores and Clark 2003, Pennington 2004). Rubin (2011) explains that the present emphasis on standardised testing in the USA as embodied by the No Child Left Behind Act is generating low levels of morale, an increase in stress and anxiety, a sense of deprofessionalisation of teaching, and teacher attrition. Such unintended outcomes have an impact on teachers' attitudes towards assessment.

Negative attitudes towards high-stakes testing might lead teachers to demonise it and disregard the fact that it can be beneficial. In fact, Taras (2005:469) argues that 'the terrors evoked by the term "assessment" have distorted its necessity, centrality and its potentially neutral position'.

Pishghadam, Adamson, Sadafian and Kan (2014:46) found that 'teachers who do not esteem assessment as a sign of school quality or an improvement tool for learning, and deem assessment negative, bad and unfair, may become exhausted, indifferent, and finally experience burnout to a higher degree'. However, when teachers are convinced that a high-stakes test is rigorously designed and has the potential to aid teaching and learning then their attitudes towards it may be positive. In her study on perceptions of English language testing in Taiwan, Wu (2008:8) found that despite some teachers' concern that external exams are the driving force behind teaching and learning, they also concede that good exams might have a positive washback effect. Teachers are more likely to perceive the introduction of external standardised assessment as motivating for students and supportive of learner autonomy if tests are deemed to be a well-designed measure of an appropriate range of knowledge and skills (Docherty, Casacuberta, Rodriguez Pazos and Canosa 2014). It seems as if the negative attitudes engendered by high-stakes testing are a result of teachers being deprived of a sense of ownership over high-stakes tests and being unconvinced of their potential to lead to quality teaching and learning. Providing teachers with ownership over high-stakes testing by encouraging them to be involved in test development might be one way of changing their attitudes towards high-stakes tests.

Teacher as examiner, examiner as teacher

Mostly characterised as negative due to the uses of high-stakes tests and the attitudes towards them, the washback effect of such tests on classroom practice is potentially strong. Nonetheless, some researchers argue that 'high-stakes tests, powerful as they are, might not be efficient agents for profound changes in an educational context' (Tsagari 2009:8). Irrespective of the level of strength, the washback effect of such tests need not always be negative and stultifying. While acknowledging that there is scant empirical evidence on the formative use of summative assessment data, Hoover and Abrams (2013) found that the majority of teachers of English and other subjects in their study used such data to change their instruction. Moreover, positive washback is more likely to ensue if tests are produced with an awareness of the learning context. According to Whitehead (2007:449), the validity of tests can be enhanced if they possess ecological validity, i.e. if they reflect teaching and learning, and students' use of the assessed content. Providing teachers with a sense of ownership by encouraging them to play an active role in high-stakes testing is likely to increase its formative potential.

Teachers' involvement in high-stakes testing can help in reducing the alienation that they sometimes experience in relation to tests that are implemented without their consultation. Gregory and Clarke (2003:72) argue that teachers must be able to engage with any assessment systems that are

about to be implemented and evaluate their strengths and weaknesses. This is crucial if they are to contribute to policy-making in relation to assessment and thus prevent the kind of centralisation of power that can damage students (Gregory and Clarke 2003:73). Teachers who are not involved in language testing may 'feel that a gap between teaching and testing is in evidence. They often feel that those who write the tests are not in touch with the realities of the classroom' (Coombe, Al-Hamly and Troudi 2009:15). Marshall (2011) discusses how the London Association for the Teaching of English acted as a platform from which teachers could take a more active role in high-stakes examinations and thus reform the assessment system by encouraging examination boards to adopt a bottom-up approach. This case study epitomises 'the growing role of the teacher as examiner, and the examiner as teacher' (Norman 2011:1,055). By being encouraged to position themselves in this way teachers are likely to feel that their judgement matters. Klenowski and Wyatt-Smith (2012:75) point out that in order for national testing programmes to improve outcomes there needs to be agreement on the idea that teachers, rather than tests, are the primary change agents. This entails foregrounding teacher judgement. The latter can serve to heighten the formative potential of high-stakes tests and it is for this reason that there should be more opportunities for teachers to play the role of examiners. Sloane and Kelly (2003:12) highlight the need for teachers to contribute to test design so that the resulting test is aligned with the curriculum and has the potential to heighten student motivation. Harlen (2005a:221) is in favour of involving teachers in public tests because through such 'involvement they develop ownership of the procedures and criteria and understand the process of assessment, including such matters as what makes an adequate sample of behaviour, as well as the goals and processes of learning'. The implication is that the knowledge and skills they develop by being involved in such high-stakes tests will feed into their own classroom practices. However, such involvement might first require bolstering their confidence in their own judgement. One way of doing this is by developing an assessment community within a school so as to increase confidence in teacher judgement amongst teachers and test users (Harlen 2005b:266). Teachers' confidence in preparing their students for high-stakes tests 'is less likely to come from pep rallies or inspirational speakers than it is from the slow, steady work of teachers working together to understand the tasks their students will face on high-stakes exams' (Reich and Bally 2010:182). Enabling teachers to position themselves as examiners empowers them to play a role in reforming high-stakes testing so that it is more equitable and more likely to enhance classroom practices.

Most probably one of the reasons for which teachers are not encouraged to be more actively involved in high-stakes testing is the perception that their assessment practices in other non-high-stakes situations are insufficiently

reliable (Brookhart 2013, Harlen 2005b). It is due to this that teacher assessment is most often pushed out of national assessment. However, there is a danger in such exclusion, particularly in relation to the validity of the assessment system. Talking about the UK context, Stobart (2001:37) argues that the validity of national curriculum assessment can only be safeguarded if there is a balance between teacher assessment and external tests. The two forms of assessment are mutually beneficial and both teachers and the assessment system stand to gain by maintaining the balance. For example, Chisholm and Wildeman (2013:98) report how in South Africa 'other forms of assessment continue to exist alongside tests and the focus is on the teacher development, infrastructural and textual resource interventions necessary to address the weaknesses revealed by tests'. An assessment system that aims to safeguard its validity while improving outcomes will seek to harness teachers' knowledge of the learning context. According to Johnson (2013:93), 'there can be little doubt that teachers represent a wealth of knowledge about students' achievements and capabilities that is indispensable in the assessment of learning progress and achievement, and which, in principle, could usefully be exploited in high-stakes examination and certification systems'. Tapping teachers' knowledge of the learning context might be carried out not only by allowing teacher assessment to complement high-stakes testing but also by providing teachers with the necessary training in order for them to contribute to the latter. For example, in the case of GCSEs, secondary school leaving examinations for 16-year-olds in the UK, the fact that teachers will still be able to receive face-to-face training focusing on the knowledge and skills they need to conduct controlled assessment is for Crisp (2013:142) an acknowledgement of the valuable role that teachers play in such assessments and the significance of providing them with adequate support. This is in line with the idea that 'teachers in general are capable of internalizing a standard accurately . . . provided they are trained in that standard' (North and Jarosz 2013:122). The solution to a lack of reliability in teachers' assessment practices is best addressed by means of training and not by barring them from participating in high-stakes testing.

Developing teachers' assessment literacy

Developing teachers' assessment literacy seems to be necessary for them to operate more effectively in an educational culture dominated by high-stakes testing. This is defined as 'the ability to design, select, interpret, and use assessment results appropriately for educational decisions' (Quilter and Gallini 2000:116). According to Gulek (2003:49), teachers 'need to be assessment literate in order to respond to the demands of the avalanche of high-stakes testing. Being assessment literate broadens one's perspective to view assessment as a dynamic process'. However, it seems as if many education

systems globally face the problem of a lack of assessment literacy among educators (Koh 2011:256). Research seems to show that teachers' assessment literacy is rather poor (Chisholm and Wildeman 2013, Earl 2003, Guskey 2004, Quilter and Gallini 2000) and this leads them to assess students in the largely ineffective way they themselves were assessed (Guskey 2004). According to Coombe et al (2009:15), 'without a higher level of teacher assessment literacy, we will be unable to help students attain higher levels of academic achievement'. Moreover, teachers' failure to understand the purpose of high-stakes testing affects their classroom practices and attitudes towards assessment (Bracey 2005, Burger and Krueger 2003, Earl 2003, Lewis 2007). Developing teachers' assessment literacy might be a means of addressing some of these problems.

Providing teachers with adequate training is crucial, especially since professional development opportunities that target teachers' assessment literacy have been associated with an improvement in student outcomes (Timperley, Wilson, Barrar and Fung 2007). Klenowski and Wyatt-Smith (2012:75) point out that high-stakes testing should serve to nurture, rather than minimise, teachers' professional abilities if tests are to contribute to student learning. Developing teachers' assessment literacy might entail having 'to divert some of the funding for test development and trialling into professional development opportunities to build teacher assessment capabilities, especially in task design and the use of achievement standards' (Klenowski and Wyatt-Smith 2012:75). According to Costigan (2002:33), teacher education programmes need to evaluate whether teacher candidates are adequately prepared for the high-stakes testing culture that is currently in existence. Teacher education programmes 'must provide opportunities for candidates to consider and discuss issues associated with high-stakes testing of their future students' (Martin, Chase, Cahill and Gregory 2011:367). This is fundamental because 'pre-service teacher education has a critical role to play in promoting assessment literacy in beginning teachers and in providing a foundation for teachers' continued learning about assessment throughout their careers' (DeLuca, Chavez and Cao 2013:123). According to Brookhart (2013:86), unless the quality of teacher judgement is addressed by means of research and practice, teachers will continue being excluded from high-stakes testing. It seems clear that training programmes targeting teachers' assessment literacy are essential, especially because they enhance teacher judgement and recognise teachers' professionalism and ability to contribute to high-stakes testing.

Training programmes would be even more effective if they consisted of teacher involvement in high-stakes testing. This would help address a problem identified by Watanabe (2011:33), who acknowledges that while promoting assessment literacy among teachers is significant, it has not yet been determined what kind of knowledge and skills need to be developed and to what extent. When teachers are involved in high-stakes testing, their own

assessment literacy is higher than when they are excluded from the process (Runté 1998). This kind of involvement serves as a valid form of professional development that has a positive impact on classroom assessment as well as on high-stakes testing. According to Tang (2010:676), 'opportunities of and support in reflection, conscious deliberation, and theorization of practice are important to bring about a more sophisticated form of professional knowledge integration'. Black, Harrison, Hodgen, Marshall and Serret (2011) show that with the right kind of strategic support, teachers' competence in summative assessment results in a positive effect on teaching and learning. However, this can only be achieved by means of extensive professional development involving hands-on work rather than just through reading material and a brief training session (Black et al 2011:463). In fact, in line with the relevant literature (McMunn, McColskey and Butler 2004, Wiliam and Thompson 2008), Koh's (2011:272) study underscores the fact that 'ongoing, sustained professional development is more powerful than short-term, one-shot professional development workshops'. Focusing on educators in Germany and Sweden, Forsberg and Wermke (2012) found that teachers' assessment literacy was mostly a product of their professional experiences and collaboration with colleagues rather than formal training. This shows that the latter needs to provide opportunities for hands-on, non-formal learning by teachers. Formal training, especially when it is theoretically oriented, needs to operate in tandem with practice-based activities.

A crucial part of any training programme is the one highlighting teachers' beliefs and attitudes in relation to assessment. Quilter and Gallini (2000:128) show that 'personal experiences with testing play an important role in understanding teachers' current attitudes toward assessment, whereas their professional training in educational measurement may play a negligible role'. This seems to imply that training needs to target not only teachers' assessment literacy but primarily their perception of assessment. In describing how two teachers implemented new assessment practices in their respective contexts, Sato, Coffey and Moorthy (2005:190) feel 'convinced that for the sustained and powerful spread of ideas, new programmes or approaches need to honour the individual teacher's priorities, visions and contexts'. This implies that in order for training to be truly effective it needs to take into account teachers' attitudes and beliefs, and draw upon their experiences and classrooms. Helping teachers to identify their beliefs about assessment is essential if they 'are to differentiate their initial ideas about assessment from the ideas they are being asked to accept, to challenge them and to integrate aspects of these new ideas into a new set of beliefs' (Vandeyar and Killen 2006:44). This is in line with the idea that a training programme needs to first identify teachers' theories, assumptions and practices and then work to improve these by encouraging participants to reflect amongst themselves (Black, Harrison, Hodgen, Marshall and Serret 2010).

Context

In 2010, MATSEC, Malta's national examination body, published a new syllabus for the Advanced English Examination that students typically sit for at the end of a two-year course at a post-16 institution like the one where we teach. This examination forms part of second language education and caters for the needs of around 600 candidates, who typically sit for the examination at the age of 18. If taken as part of a Matriculation Certificate, candidates would usually aspire to further their studies at the country's sole university. Most undergraduate courses that consider Advanced English to be one of their entry requirements specify that applicants need to have a pass at grade C or better. Even though it has not been officially aligned to the CEFR, this would be equivalent to at least C1.2 level, i.e. the upper end of effective operational proficiency. The 2010 syllabus contained a brief outline of a speaking component that had not featured in previous syllabuses. It also specified that the first sitting of this speaking examination would take place in May 2013, allowing adequate time for post-16 institutions to start developing their students' speaking skills.

The introduction of this speaking component served to address a lacuna in relation to the testing of candidates' speaking skills, a lacuna that was allowing candidates to be awarded a qualification testifying to their high level of proficiency in English without ever needing to demonstrate evidence of spoken fluency. The revised syllabus meant that suddenly it was considered 'desirable that candidates studying English at Advanced level demonstrate an evolved proficiency in speaking and listening skills' (MATSEC 2010:6). The speaking component was intended to act 'as a measure of the candidates' ability to speak and converse in English' (MATSEC 2010:6). The Advanced English Examination was finally catering for the assessment of oracy.

Together with our colleagues within the English department of Malta's largest post-16 school, we welcomed the inclusion of a speaking component in the Advanced English Examination, however, we were somewhat disappointed by the lack of detail in the syllabus's description of this component. We felt that as teachers we were not sufficiently confident as to what was expected of our students in each part of the speaking examination and that more detailed specifications were required in order for us to know what they were going to be assessed on. The speaking examination was meant to be developed by MATSEC to serve the needs of the national cohort but it was immediately clear to us that there would not be any further elaboration beyond the syllabus description unless we took the lead to develop the examination by first of all writing a manual for the benefit of all stakeholders. For a number of years we had worked as examiners for the *International English Language Testing System* (*IELTS*), *Cambridge English: Advanced*, also known as *Certificate in Advanced English* (*CAE*), and *Cambridge*

English: Proficiency, also known as *Certificate of Proficiency in English* (*CPE*). This familiarity with international examinations made us realise that what both teachers and candidates truly needed was a comprehensive test manual that elaborated on each part of the new speaking component and provided detailed information on content, structure, timing, techniques, criterial levels of performance, and scoring procedures. MATSEC had never produced thorough examination manuals in accompaniment to syllabuses. Hence, together with a small group of colleagues, we took the initiative to create such a manual. Our decision was supported by the Head of Department, who also formed part of the team. We reasoned that by doing so we would be able to set a standard for the examination that would be extremely hard to reject. Moreover, such a manual would improve our ability to fulfil our roles more effectively and be of benefit to our students and other examination candidates. Despite the fact that we had not been commissioned by MATSEC to do this work, it had been informally told that it was being carried out.

After reviewing a number of manuals for a range of international speaking examinations as well as assessment textbooks by Hughes (2003), Fulcher (2003) and Luoma (2004) amongst others, we realised that there were many elements we wanted to incorporate into the Advanced English Speaking Examination in order for it to be a more reliable, valid, and equitable example of a high-stakes test. We wanted to produce a speaking examination that the relevant stakeholders could value. In the process of discussing the different decisions that needed to be taken in order to improve the syllabus description of this examination, we became aware that our own assessment literacy, attitudes and beliefs were being developed by the very act of reflecting on what suited the needs of the colleagues we worked with and the hundreds of students we taught.

The following sections describe the stages we followed in developing the speaking examination, starting with the writing of the specifications and their eventual incorporation into a manual, moving on to the writing of items, moderation and trialling, followed by the development of a rating scale, and finally ending with examiner training.

Examination specifications

Our first challenge was to write a comprehensive set of specifications for this speaking component while keeping in mind the syllabus outline. Given the official nature of the syllabus we could not avoid working within the parameters set by its description of the component. The syllabus specified that the speaking examination was to carry 6% of the global mark and to last a total of 15 minutes. Moreover, the syllabus described this component as being made up of three parts:

> Part 1 is a guided examiner-to-candidate conversation.
> Part 2 is a guided examiner-to-candidate conversation.
> Part 3 is a candidate-to-examiner 'long turn' (MATSEC 2010:6–7).

An appendix to the syllabus gave an example of the kind of task/s that the candidate would have to complete in each part. It also indicated the approximate amount of time that each part should take as well as the number of marks allotted to each one.

In deciding to write this examination's specifications we were subscribing to the idea that 'the greater the detail in the specification of content, the more valid the test is likely to be' (Hughes 2003:116). We decided to write a thorough explication of each part by first presenting its aims and content, and then carefully listing the procedures to be used by examiners. We knew that this information would serve as the backbone of the examination manual we wanted to present to MATSEC at the end of the process.

Aims and content

Our description of the aims and content not only specified what kind of tasks candidates would have to complete in each part but also explained what they were expected to achieve in doing so. For example, the syllabus specified that Part 1 was 'an informal interview intended as a conversation starter, where the examiner will ask basic questions about topics such as work, study, leisure and career plans' (MATSEC 2010:6). In our gloss we thought that it would be more helpful to inform stakeholders that the purpose behind Part 1 was to assess candidates' ability to give basic information about themselves and express general views as well as specific details on familiar topics. This was based on our familiarity with international speaking examinations and with recommendations made by the literature on assessing speaking. We also considered it expedient to indicate that the examiner's questions could focus on past, present and future situations, and that they were not meant to be specifically challenging in terms of language and content. In this way we sought to underscore the fact that Part 1 was intended to enable candidates to talk about what was highly familiar to them before being expected to engage in more demanding tasks in the following two parts of the examination.

Examination procedures

Our elaboration on the procedures was meant to be as exhaustive as possible so that examiners would behave in a standardised fashion when conducting the examination. In this way we sought to enhance the examination's reliability. This entailed scripting most of what the examiners were required to say in conducting the examination and providing them with information

about what they were expected to do in case candidates were unable to sustain a particular turn or gave overlong responses. It also necessitated being highly specific about the structure and timing of each part of the examination. The syllabus used the word 'about' to describe how long each part should take and this provided us with some crucial leeway when specifying the exact amount of time that each part was meant to take. For example, according to the syllabus, Part 2 was meant to take 'about four minutes' and to involve 'a conversation initiated by the interlocutor, based on a prompt such as a photograph or other image that is presented to the candidate at this point in the interview' (MATSEC 2010:6). After the candidate 'briefly' describes the picture, 'The examiner will then follow one set of questions from a number of options available' (MATSEC 2010:9). In writing the procedures for Part 2 we wanted to confirm for examiners and candidates how long each one of these two stages should actually take. Based on our experience, we also considered it fair that candidates should have some time in which to study the visual prompt before describing it. Hence we agreed that the examiner should first present the candidate with the prompt and then provide them with 30 seconds in which to look at it before asking them to describe it. We specified that the description should not last longer than 1 minute and that this was to be followed by a two-way exchange between the examiner and the candidate lasting no longer than 3 minutes. By being so attentive to timing we felt that candidates preparing for this examination would know exactly what was expected of them in each part. We considered this important given our aim to achieve accountability by means of the examination manual (McNamara 2000).

Examination manual

The examination manual we wrote was not conceived to be a monolithic document. Fulcher (2003:116) claims that 'specifications are dynamic, evolving, documents that should be related to the process of test design, piloting and revision'. The specifications we wrote for the Advanced English Speaking Examination are a product of reflection and discussion, trialling, and research data. Moreover, we always intended them to be open to regular reevaluation. This resonates with Luoma's (2004:116) idea that the act of writing specifications is educational for novice test developers given that it facilitates the process of 'making concrete connections between the theory and practice of oral assessment in their own context, through their own data'. We realised that the more we would learn about the examination, the more adept we would become at improving its specifications.

Two versions of the examination manual were written, one for the needs of teachers and candidates and another one for the needs of examiners. The former version sought to reassure candidates that they would not

experience any surprises upon sitting for the examination. We considered this to be fundamental given that 'the degree of a test taker's familiarity with the demands of a particular test may affect the way the task is dealt with' (Weir 2005:54). The examiners' version of the manual guided them as to what they should do in each part and when they should do it. This involved instructing examiners of the exact time when they should provide candidates with printed prompts and when to collect them as well as outlining what kind of assistance should be provided to candidates. For example, the syllabus specified that Part 3 should last 'about 3 minutes' and in it a candidate should engage in a 'presentation expressed as a long turn . . . based on a question selected by the candidate from a list of five presented to her/him some minutes before entering the interview room' (MATSEC 2010:6–7). We decided that if candidates were to be given 10 minutes before the beginning of the examination to prepare a 3-minute presentation they would need some time during the actual examination to go over the main points of the presentation and make notes. Hence we agreed that it was only reasonable to provide candidates with a maximum of 2 minutes at the start of Part 3 in which they could gather their thoughts and jot down any important points on a sheet of blank paper given to them by the examiner. Candidates were to be allowed to refer to these notes when delivering their presentation but they were to be handed to the examiner at the end of the examination in order to avoid cheating. On the basis of our experience and the trialling of sample test materials we felt that our students would most likely consider Part 3 to be the toughest, so by introducing these procedures we were providing candidates with an opportunity to be fairly assessed on their speaking skills.

Writing items, moderation and trialling

After having constructed a comprehensive set of specifications for the Advanced English Speaking Examination, we considered it worthwhile to write specimen test materials in line with these specifications (Hughes 2003:63). This was particularly important given that for inexperienced test developers 'writing specifications together with the first versions of the tasks and scales will help them avoid some problems with test use' (Luoma 2004:115). We formed three groups and each one focused on creating materials for a specific part of the examination. In writing these materials we sought to imagine how our students would interpret the wording of the different tasks in each part of the examination. In this way we attempted to preempt any misinterpretations and ascertain that the tasks were truly testing what we intended them to test, hence ensuring validity. Subsequently, we exchanged materials so that we could moderate each other's work. In addition, a colleague who was external to the whole process contributed to this end.

In line with its purpose (Hughes 2003:63), moderation allowed us to identify a number of flaws in the tasks we had created and make the necessary adjustments.

We trialled the specimen examination materials with a group of students who were very similar in terms of age and educational level to the examination's eventual candidates. We each conducted five to eight mock sessions and we also observed one another on several occasions. Before every session we provided the student with as much information about the content and structure of each part so as to approximate the level of familiarity that a typical candidate would have after adequate preparation for the examination. Whilst conducting these sessions we made a note of any problems we encountered and if colleagues were acting as observers they did the same. At the end of every session we asked for detailed feedback from the student so that we could factor in the students' point of view of the examination we had designed. Moreover, a number of colleagues working at other post-16 schools were asked to trial the materials we had developed with a sample of their own students. Trialling allowed us to make further changes to the examination materials as well as to tweak the procedures we had devised. It is for this reason that trialling is described as a 'critical phase of the work' (Fulcher 2003:118) involved in test development. Trialling 'ensures that there is sufficient time available for candidates to produce a situationally and interactionally authentic spoken contribution' (Galaczi and ffrench 2011:137). For example, it was only after the materials had been trialled in four different schools that we took the decision to extend the preparation time in Part 3 to 2 minutes. Initially, we had specified that this should not be longer than 1 minute, but after trialling it became clear that this was insufficient for most students. This kind of trialling was highly useful as it enabled us to improve the design of the examination. However, we realised that further trialling would be necessary once we had developed a rating scale and calibrated it. The latter process was significant given that MATSEC had not specified any criteria in terms of the level of proficiency expected of candidates.

Rating scale calibration

Given that there was no indication on the part of MATSEC of what kind of instrument was going to be used in order to assess candidates' speaking skills we agreed that the best way of doing this was via an analytic rating scale. The four assessment criteria we opted to base our rating scale on were: fluency and coherence; pronunciation; vocabulary; and grammar. Like Weir (2005:191), we considered it 'useful if the criteria employed in the assessment of language production on tasks could be related in a principled way to the criteria for the teaching of a skill'. We wanted the examination to have a

positive washback effect on certain aspects of speaking deemed as a priority in the classroom context.

The main hurdle we faced was that the syllabus did not only specify a global mark (i.e. 18 marks) for the speaking component but also prescribed the total number of marks for each part (i.e. Part 1: 4 marks; Part 2: 6 marks; Part 3: 8 marks). This meant that we could not easily use a system of bands as in the *IELTS* test. Hence, as shown in Table 1, we decided to subdivide each set of marks into three groups and create a separate descriptor for each group in terms of each one of the four assessment criteria. This meant that we had to write a total of 36 descriptors. However, some of the descriptors for certain criteria could be used for more than one part of the examination.

Table 1 Rating scale example

Part 1			
Marks	1–2	3	4
Fluency & Coherence	Descriptor	Descriptor	Descriptor

The writing of the descriptors was a lengthy process that underwent many redrafts by different members of the team. We were only satisfied with the rating scale once trialling and our professional experience led us to feel convinced that the three descriptor levels for each assessment criterion would enable examiners to discriminate amongst different candidates depending on their level of speaking proficiency in a particular area.

In order to calibrate the scale we decided to video record a series of 50 mock examinations with students who had a very similar profile to the candidates who would eventually be sitting for the Advanced English Speaking Examination. In order to do this we trained a group of three colleagues in terms of the content and procedures we had devised; for the sake of standardisation we used the same set of specimen examination materials we had developed for the manual. Before we started filming the sessions we needed for calibration purposes, we ensured that these three 'examiners' had mastered the procedures and materials through a number of non-recorded sessions. The 50 videos provided us with samples of performance covering the entire range of the scale (Hughes 2003). Subsequently, we assigned each one of these samples to a relevant point on the scale in terms of each criterion and for each part. However, we chose to focus on calibrating the descriptors of each part in turn so as to facilitate our understanding and effective use of these descriptors. For example, we would watch Part 1 of a session and then each one of us would assign marks for every criterion. After assigning marks individually we would discuss amongst ourselves the reasons why a student was assigned that mark for that particular criterion in that specific part of

the examination. The initial challenge was to reach a consensus but once we grew familiar with how to interpret the descriptors in a uniform manner, our scoring became highly standardised. The videos we used to calibrate the rating scale eventually became a crucial part of the training of examiners given that they constituted reference points for the different descriptors.

Examiner training

Once we had finalised the rating scale and the elucidation of the examination's content and procedures, we contacted MATSEC to let it know of our work. Despite the fact that we were not commissioned by MATSEC to do this work, we felt confident that it would appreciate our efforts and seek to implement them. This, in fact, happened and we were invited to present the draft of our manual to all the English teachers working in post-16 institutions in Malta. These teachers were encouraged to test the specimen materials with their students and to provide us with feedback; on the basis of this feedback we made some slight changes to the manual.

Subsequently, MATSEC agreed with our suggestion to organise a training course for prospective speaking examiners and we were asked to run this course. According to Luoma (2004:177–178), 'the use of rater training . . . means that the developers recognise the impossibility of giving comparable ratings without training, and they take steps to ensure comparability because they consider it important'. A group of 18 trainee examiners made up mostly of teachers working in post-16 schools attended the 9-hour training course and by means of it we sought to consolidate their understanding of the different parts of the speaking examination, the assessment criteria, and the examination procedures. We also aimed to provide them with plenty of practice in conducting the examination and in using the rating scale to assess a candidate's performance. For the purposes of the training course we created a list of FAQs that future examiners could refer to when they had any queries about a number of common issues that we had encountered whilst carrying out the mock examinations. Our intention was to add to these FAQs after each sitting of the examination.

After adequate numbers of hands-on activities intended to familiarise the trainee examiners with structure, rubrics and timing, they practised using the rating scale by means of the samples of performance we had recorded. This was in line with Hughes's (2003) suggestion to use calibrated videos when training speaking examiners. Each time the trainees were shown a video clip, they were asked to write down the marks they had assigned to the 'candidate' according to the rating scale descriptors; then a discussion followed. The rating forms used for this purpose were kept as a record of the trainees' performance. Fulcher (2003:145) points out that 'the process of rater training is designed to "socialize" raters into a common understanding of the scale

descriptors, and train them to apply these consistently in operational speaking tests'. Achieving a satisfactory level of standardisation in scoring was a lengthy and challenging process but eventually we were pleased that a fair number of trainees were using the rating scale appropriately, leading to a high degree of intra-rater and inter-rater reliability. By the end of the course around half of the trainees were certified as examiners. Given our awareness of training attrition, we envisaged the course to be part of 'a cyclical, iterative process which goes beyond the initial standardization phase' (Taylor and Galaczi 2011:213). Hence we specified that these examiners needed to be recertified every year. As test developers and trainers, we joined this group of examiners in conducting the sessions forming part of the first sitting of the Advanced English Speaking Examination. Our participation in assessing actual candidates was the culmination of three years' involvement in different stages of this examination.

Conclusion

When a high-stakes test is introduced at a national level it is important that the teachers who are going to be affected by this test are provided with the right kind of support so that they may understand the purposes and procedures governing the test. Yip and Cheung (2005:161) affirm that 'to empower teachers, they should be given more opportunities to develop the knowledge and skills necessary for implementing the innovations, through the provision of supporting materials and the organisation of training workshops and courses or experience-sharing sessions'. Involving teachers in the design and implementation of the test is a highly effective way of ensuring such empowerment. It can be argued that developing educators' professional judgement through such involvement is crucial given that the ideal of scientific measurement is impossible to attain (Yorke 2011). Our experience has shown us that the involvement of teachers in high-stakes testing leads to an increased level of assessment literacy, the cultivation of positive beliefs and attitudes in relation to assessment, the bolstering of confidence in teachers' judgement, and more equitable examinations. Further research might indicate whether such implications resonate with the experiences of educators in other contexts.

Acknowledgements

The design and implementation of the Advanced English Speaking Examination described in this paper was conducted together with our colleagues Clyde Borg, Andrew Farrugia, Joseph Gerardi, and Odette Vassallo. We would like to thank the British Council for having recognised the merits of our project by awarding us the 2014 Innovation in Assessment Prize.

References

Allen, A (2012) Cultivating the myopic learner: the shared project of high-stakes and low-stakes assessment, *British Journal of Sociology of Education* 33 (5), 641–659.

Assaf, L C (2008) Professional identity of a reading teacher: responding to high-stakes testing pressures, *Teachers and Teaching: Theory and Practice* 14 (3), 239–252.

Atkin, J M (2007) Swimming upstream: relying on teachers' summative assessments, *Measurement: Interdisciplinary Research and Perspectives* 5 (1), 54–57.

Au, W W (2008) Devising inequality: a Bernsteinian analysis of high-stakes testing and social reproduction in education, *British Journal of Sociology of Education* 29 (6), 639–651.

Black, P, Harrison C, Hodgen J, Marshall, B and Serret, N (2010) Validity in teachers' summative assessments, *Assessment in Education: Principles, Policy & Practice* 17 (2), 215–232.

Black, P, Harrison, C, Hodgen, J, Marshall, B and Serret, N (2011) Can teachers' summative assessments produce dependable results and also enhance classroom learning? *Assessment in Education: Principles, Policy & Practice* 18 (4), 451–469.

Bracey, G W (2005) The 15th Bracey report on the condition of public education, *Phi Delta Kappan* 87 (2), 138–153.

Brookhart, S M (2013) The use of teacher judgement for summative assessment in the USA, *Assessment in Education: Principles, Policy & Practice* 20 (1), 69–90.

Burger, J M and Krueger, M (2003) A balanced approach to high-stakes achievement testing: an analysis of the literature with policy implications, *International Electronic Journal for Leadership in Learning* 7 (4), available online: www.ucalgary.ca/iejll/burger_krueger

Chisholm, L and Wildeman, R (2013) The politics of testing in South Africa, *Journal of Curriculum Studies* 45 (1), 89–100.

Christopher, N M (2009) Interrelation between environmental factors and language assessment in Nigeria, *Research Notes* 35, 10–15.

Coombe, C, Al-Hamly, M and Troudi, S (2009) Foreign and second language teacher assessment literacy: issues, challenges and recommendations, *Research Notes* 38, 14–18.

Costigan, A T (2002) Teaching the culture of high stakes testing: listening to new teachers, *Action in Teacher Education* 23 (4), 28–34.

Crisp, V (2013) Criteria, comparison and past experiences: how do teachers make judgements when marking coursework? *Assessment in Education: Principles, Policy & Practice* 20 (1), 127–144.

Crocco, M S and Costigan, A T (2006) High-stakes teaching: what's at stake for teachers (and students) in the age of accountability, *The New Educator* 2 (1), 1–13.

DeLuca, C, Chavez, T and Cao, C (2013) Establishing a foundation for valid teacher judgement on student learning: the role of pre-service assessment education, *Assessment in Education: Principles, Policy & Practice* 20 (1), 107–126.

Docherty, C, Casacuberta, G G, Rodriguez Pazos, G and Canosa, P (2014) Investigating the impact of assessment in a single-sex educational setting in Spain, *Research Notes* 58, 3–15.

Earl, L M (2003) *Assessment as Learning: Using Classroom Assessment to Maximize Student Learning*, Thousand Oaks: Corwin.
Flores, B B and Clark, E R (2003) Texas voices speak out about high-stakes testing: preservice teachers, teachers, and students, *Current Issues in Education* 6 (3), available online: cie.asu.edu/volume6/number3/
Forsberg, E and Wermke, W (2012) Knowledge sources and autonomy: German and Swedish teachers' continuing professional development of assessment knowledge, *Professional Development in Education* 38 (5), 741–758.
Fulcher, G (2003) *Testing Second Language Speaking*, Harlow: Pearson Education.
Galaczi, E and ffrench, A (2011) Context validity, in Taylor, L (Ed) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press, 112–170.
Grant, C A (2004) Oppression, privilege, and high-stakes testing, *Multicultural Perspectives* 6 (1), 3–11.
Gregory, K and Clarke, M (2003) High-stakes assessment in England and Singapore, *Theory Into Practice* 42 (1), 66–74.
Gulek, C (2003) Preparing for high-stakes testing, *Theory Into Practice* 42 (1), 42–50.
Guskey, T (2004) Zero alternatives, *Principal Leadership* 5 (2), 49–53.
Harlen, W (2005a) Teachers' summative practices and assessment for learning: tensions and synergies, *Curriculum Journal* 16 (2), 207–223.
Harlen, W (2005b) Trusting teachers' judgement: research evidence of the reliability and validity of teachers' assessment used for summative purposes, *Research Papers in Education* 20 (3), 245–270.
Hoffman, J, Assaf, L C and Paris, S (2001) High stakes testing in reading: today in Texas, tomorrow? *The Reading Teacher* 54, 482–499.
Hoover, N R and Abrams, L M (2013) Teachers' instructional use of summative student assessment data, *Applied Measurement in Education* 26 (3), 219–231.
Hughes, A (2003) *Testing for Language Teachers* (2nd edition), Cambridge: Cambridge University Press.
Johnson, S (2013) On the reliability of high-stakes teacher assessment, *Research Papers in Education* 28 (1), 91–105.
Jones, B D (2007) The unintended outcomes of high-stakes testing, *Journal of Applied School Psychology* 23 (2), 65–86.
Klenowski, V and Wyatt-Smith, C (2012) The impact of high stakes testing: the Australian story, *Assessment in Education: Principles, Policy & Practice* 19 (1), 65–79.
Koh, K H (2011) Improving teachers' assessment literacy through professional development, *Teaching Education* 22 (3), 255–276.
Lewis, A C (2007) How well has NCLB worked? How do we get the revisions we want? *Phi Delta Kappan* 88 (5), 353–358.
Lewkowicz, J and Zawadowska-Kittel, E (2008) Impact of the new school-leaving exam of English in Poland, *Research Notes* 34, 27–31.
Luoma, S (2004) *Assessing Speaking*, Cambridge: Cambridge University Press.
Marshall, B (2011) *Testing English: Formative and Summative Approaches to English Assessment*, London: Continuum.
Martin, S D, Chase, M, Cahill, M A and Gregory, A E (2011) Minding the gate: challenges of high-stakes assessment and literacy teacher education, *The New Educator* 7 (4), 352–370.

MATSEC Examinations Board (2010) *Advanced Matriculation Syllabus English 2013*, available online: www.um.edu.mt/__data/assets/pdf_file/0015/108024/AM10.pdf

McMunn, N, McColskey, W and Butler, S (2004) Building teacher capacity in classroom assessment to improve student learning, *International Journal of Educational Policy, Research, & Practice* 4 (4), 25–48.

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

Nichols, S L (2007) High-stakes testing, *Journal of Applied School Psychology* 23 (2), 47–64.

Nichols, S L and Berliner, D C (2007) *Collateral Damage: How High-stakes Testing Undermines Education*, Cambridge: Harvard Education Press.

Norman, P (2011) Testing English: formative and summative approaches to English assessment, *British Educational Research Journal* 37 (6), 1,055–1,057.

North, B and Jarosz, E (2013) Implementing the CEFR in teacher-based assessment: approaches and challenges, in Galaczi, E D and Weir, C J (Eds) *Exploring Language Frameworks*, *Proceedings of the ALTE Kraków Conference, July 2011*, Studies in Language Testing volume 36, Cambridge: UCLES/Cambridge University Press, 118–134.

Pennington, J L (2004) Teaching interrupted: the effect of high-stakes testing on literacy instruction in a Texas elementary school, in Boyd, F B and Brock, C H (Eds) *Multicultural and Multilingual Literacy and Language*, New York: Guilford Press, 241–261.

Pishghadam, R, Adamson, B, Sadafian, S S and Kan, F L F (2014) Conceptions of assessment and teacher burnout, *Assessment in Education: Principles, Policy & Practice* 21 (1), 34–51.

Quilter, S M and Gallini, J K (2000) Teachers' assessment literacy and attitudes, *The Teacher Educator* 36 (2), 115–131.

Reich, G A and Bally, D (2010) Get smart: facing high-stakes testing together, *The Social Studies* 101 (4), 179–184.

Rubin, D I (2011) The disheartened teacher: living in the age of standardisation, high-stakes assessments, and No Child Left Behind (NCLB), *Changing English: Studies in Culture and Education* 18 (4), 407–416.

Runté, R (1998) The impact of centralized examinations on teacher professionalism, *Canadian Journal of Education* 23 (2), 166–181.

Sasanguie, D, Elen, J, Clarebout, G, Van den Noortgate, W, Vandenabeele, J and De Fraine, B (2011) Disentangling instructional roles: the case of teaching and summative assessment, *Studies in Higher Education* 36 (8), 897–910.

Sato, M, Coffey, J and Moorthy, S (2005) Two teachers making assessment for learning their own, *Curriculum Journal* 16 (2), 177–191.

Sloane, F C and Kelly, A E (2003) Issues in high-stakes testing programs, *Theory Into Practice* 42 (1), 12–17.

Stobart, G (2001) The validity of National Curriculum assessment, *British Journal of Educational Studies* 49 (1), 26–39.

Tang, S Y F (2010) Teachers' professional knowledge construction in Assessment for Learning, *Teachers and Teaching: Theory and Practice* 16 (6), 665–678.

Taras, M (2005) Assessment: summative and formative: some theoretical reflections, *British Journal of Educational Studies* 53 (4), 466–478.

Taylor, L (2005) Washback and impact: the view from Cambridge ESOL, *Research Notes* 20, 2–3.

Taylor, L and Galaczi, E (2011) Scoring validity, in Taylor, L (Ed) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*,

Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press, 171–233.

Timperley, H, Wilson, A, Barrar, H and Fung, I (2007) *Teacher Professional Learning and Development: Best Evidence Synthesis*, Wellington: Ministry of Education, available online: www.educationcounts.govt.nz/publications/series/2515/15341

Tsagari, D (2009) Revisiting the concept of test washback: investigating FCE in Greek language schools, *Research Notes* 35, 5–10.

Vandeyar, S and Killen, R (2006) Beliefs and attitudes about assessment of a sample of student teachers in South Africa, *Africa Education Review* 3 (1–2), 3–47.

Watanabe, Y (2011) Teaching a course in assessment literacy to test takers: its rationale, procedure, content and effectiveness, *Research Notes* 46, 29–34.

Weir, C J (2005) *Language Testing and Validation: An Evidence-Based Approach*, Basingstoke: Palgrave Macmillan.

Whitehead, D (2007) Literacy assessment practices: moving from standardised to ecologically valid assessments in secondary schools, *Language and Education* 21 (5), 434–452.

Wiliam, D and Thompson, M (2008) Integrating assessment with learning: what will it take to make it work? in Dwyer, C A (Ed) *The Future of Assessment: Shaping Teaching and Learning*, New York: Lawrence Erlbaum Associates, 53–82.

Wu, J (2008) Views of Taiwanese students and teachers on English language testing, *Research Notes* 34, 6–9.

Yip, D Y and Cheung, D (2005) Teachers' concerns on school-based assessment of practical work, *Journal of Biological Education* 39 (4), 156–162.

Yorke, M (2011) Summative assessment: dealing with the 'measurement fallacy', *Studies in Higher Education* 36 (3), 251–273.