English Language Education

John Read *Editor*

Post-admission Language Assessment of University Students

Springer

English Language Education

Volume 6

More information about this series at http://www.springer.com/series/11558

John Read
Editor

Post-admission Language Assessment of University Students

Springer

Editor
John Read
School of Cultures, Languages and Linguistics
University of Auckland
Auckland, New Zealand

ISSN 2213-6967 ISSN 2213-6975 (electronic)
English Language Education
ISBN 978-3-319-39190-8 ISBN 978-3-319-39192-2 (eBook)
DOI 10.1007/978-3-319-39192-2

Library of Congress Control Number: 2016948219

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Preface

This volume grew out of two conference events that I organised in 2013 and 2014. The first was a symposium at the Language Testing Research Colloquium in Seoul, South Korea, in July 2013 with the title "Exploring the diagnostic potential of post-admission language assessments in English-medium universities". The other event was a colloquium entitled "Exploring post-admission language assessments in universities internationally" at the Annual Conference of the American Association for Applied Linguistics (AAAL) in Portland, Oregon, USA, in March 2014. The AAAL symposium attracted the attention of the Springer commissioning editor, Jolanda Voogt, who invited me to submit a proposal for an edited volume of the papers presented at one conference or the other. In order to expand the scope of the book, I invited Edward Li and Avasha Rimbiritch, who were not among the original presenters, to prepare additional chapters. Several of the chapters acquired an extra author along the way to provide specialist expertise on some aspects of the content.

I want to express my great appreciation first to the authors for the rich and stimulating content of their papers. On a more practical level, they generally met their deadlines to ensure that the book would appear in a timely manner and they willingly undertook the necessary revisions of their original submissions. Whatever my virtues as an editor, I found that as an author I tended to trail behind the others in completing my substantive contributions to the volume.

At Springer, I am grateful to Jolanda Voogt for seeing the potential of this topic for a published volume and encouraging us to develop it. Helen van der Stelt has been a most efficient editorial assistant and a pleasure to work with. I would also like to thank the series editors, Chris Davison and Andy Gao, for their ongoing support and encouragement. In addition, two anonymous reviewers of the draft manuscript gave positive feedback and very useful suggestions for revisions.

The concerns addressed in this book are of increasing importance to English-medium universities and other institutions which are admitting students from diverse language backgrounds. We hope that these contributions will help to clarify the issues and offer a range of concrete solutions to the challenge of ensuring that students' language and literacy needs are being met.

Auckland, New Zealand John Read
April 2016

Contents

Contributors

Natasha Artemeva School of Linguistics and Language Studies, Carleton University, Ottawa, Canada

Vahid Aryadoust National Institute of Education, Nanyang Technological University, Singapore, Republic of Singapore

Christopher Denman Humanities Research Center, Sultan Qaboos University, Muscat, Oman

Cathie Elder Language Testing Research Centre, University of Melbourne, Melbourne, Australia

Janna Fox School of Linguistics and Language Studies, Carleton University, Ottawa, Canada

April Ginther Department of English, Purdue University, West Lafayette, IN, USA

John Haggerty Department of Language and Literacy Education, University of British Columbia, Vancouver, Canada

Michael Harrington School of Languages and Cultures, University of Queensland, Brisbane, Australia

Nancy L. Kauper Oral English Proficiency Program, Purdue University, West Lafayette, IN, USA

Ute Knoch Language Testing Research Centre, University of Melbourne, Melbourne, Australia

Edward Li Center for Language Education, The Hong Kong University of Science and Technology, Hong Kong, China

Sally O'Hagan Language Testing Research Centre, University of Melbourne, Melbourne, Australia

Rebecca Patterson Office of the Dean: Humanities, University of the Free State, Bloemfontein, South Africa

Anna Pot Office of the Dean: Humanities, University of the Free State, Bloemfontein, South Africa

Avasha Rambiritch Unit for Academic Literacy, University of Pretoria, Pretoria, South Africa

Michelle Raquel Centre for Applied English Studies, University of Hong Kong, Hong Kong, China

John Read School of Cultures, Languages and Linguistics, University of Auckland, Auckland, New Zealand

Thomas Roche SCU College, Southern Cross University, Lismore, NSW, Australia

Yogesh Sinha Department of English Language Teaching, Sohar University, Al Sohar, Oman

Suthathip Ploy Thirakunkovit English Department, Mahidol University, Bangkok, Thailand

Alan Urmston English Language Centre, Hong Kong Polytechnic University, Hong Kong, China

Janet von Randow Diagnostic English Language Needs Assessment, University of Auckland, Auckland, New Zealand

Albert Weideman Office of the Dean: Humanities, University of the Free State, Bloemfontein, South Africa

Xun Yan Department of Linguistics, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, USA

Part I
Introduction

Chapter 1
Some Key Issues in Post-Admission Language Assessment

John Read

Abstract This chapter introduces the volume by briefly outlining trends in English-medium higher education internationally, but with particular reference to post-entry language assessment (PELA) in Australian universities. The key features of a PELA are described, in contrast to a placement test and an international proficiency test. There is an overview of each of the other chapters in the book, providing appropriate background information on the societies and education systems represented: Australia, Canada, Hong Kong, the USA, New Zealand, Oman and South Africa. This is followed by a discussion of three themes running through several chapters. The first is how to validate post-admission language assessments; the second is the desirability of obtaining feedback from the test-takers; and the third is the extent to which a PELA is diagnostic in nature.

Keywords English-medium higher education • Post-entry language assessment (PELA) • Post-admission language assessment • Validation • Test-taker feedback • Language diagnosis

1 Overview of the Topic

In a globalised world universities in the major English-speaking countries have for some time been facing the challenges posed by student populations which have become linguistically very diverse. There are several trends which account for this diversity (see Murray 2016, for a comprehensive account). One is certainly the vigorous recruitment of international students, on whose tuition fees many university budgets are now critically dependent. In addition, the domestic population in these countries is much more multilingual as a result of immigration inflows, including many parents who migrate specifically to seek better educational opportunities for their children. A third influence is the adoption of national policies to broaden

J. Read (✉)
School of Cultures, Languages and Linguistics, University of Auckland, Auckland, New Zealand
e-mail: ja.read@auckland.ac.nz

J. Read (ed.), *Post-admission Language Assessment of University Students*, English Language Education 6, DOI 10.1007/978-3-319-39192-2_1

participation in higher education by underrepresented groups in the society, such as ethnic minorities or those from low-income backgrounds.

A complementary development is the growth in the number of universities in other countries where the instruction is partly or wholly through the medium of English. This reflects the status of English as the dominant language of international communication (Crystal 2003; Jenkins 2007; Phillipson 2009), and in academia in particular. Given the worldwide reach of the British Empire in the eighteenth and nineteenth centuries, English-medium education is not a new phenomenon, at least for colonial elites, but it has spread more widely in recent decades. Phillipson (2009) gives a useful overview of countries outside the traditionally English-speaking ones where English-medium universities are to be found:

1. Of mature vintage in some *former* '*colonies*' (South Africa, the Indian subcontinent, the Philippines)
2. Younger in *other postcolonial contexts* (Brunei Darussalam, Hong Kong, Singapore, South Pacific)
3. Well established for some *elites* (Turkey, Egypt)
4. Recent in parts of the *Arab world* (Saudi Arabia, United Arab Emirates)
5. Even more recent in *continental Europe*. (2009, p. 200; Italics in original)

Other nations like China, South Korea and Japan can certainly be added to the list.

In all these countries, whether "English-speaking" or not, it cannot be assumed that students entering the university are adequately prepared to cope with the language and literacy demands of degree studies through the medium of English. There are obviously a number of ways in which institutions can respond to this challenge, but the one which is the focus of this book is the introduction of a language assessment to be administered to students entering the university, in order to identify those who have significant academic language needs (to the extent that they are at risk of failure or not achieving their academic potential), and to guide or direct such students to appropriate forms of academic language development as they pursue their studies.

In Australia, the term "post-entry language assessment", or PELA, has come to be used for this kind of assessment programme. Australia is one of the major recipient countries of international students, as a result of the energetic recruiting strategies of its marketing organisation, IDP Education, and individual tertiary institutions throughout the country. IDP is also the Australian partner in the International English Language Testing System (IELTS), which is the preferred English proficiency test in Australia and has grown to be the market leader worldwide. Although international students routinely need to achieve a minimum IELTS score for entry, there have been ongoing concerns about the adequacy of their English proficiency to cope with the language demands of their degree programmes. Matters came to a head with the publication of an article by Birrell (2006), an Australian academic specialising in immigration research, who produced evidence that students were graduating with degrees in accounting and information technology, yet were unable to obtain the minimum score of 6.0 on IELTS needed for permanent residence and

employment in Australia. This score (or 6.5 in many cases) is the standard requirement for direct admission to undergraduate degree programmes, but the problem was that many students were following alternative pathways into the universities which allowed them to enter the country originally with much lower test scores, and they had not been re-tested at the time they were accepted for degree-level study.

Media coverage of Birrell's work generated a large amount of public debate about English language standards in Australian universities. A national symposium (AEI 2007) organised by a federal government agency was held in Canberra to address the issues and this led to a project by the Australian Universities Quality Agency (AUQA) to develop the Good Practice Principles for English Language Proficiency for International Students in Australian Universities (AUQA 2009). The principles have been influential in prompting tertiary institutions to review their provisions for supporting international students and have been incorporated into the regular cycle of academic audits conducted by the AUQA and its successor, the Tertiary Education Quality and Standards Agency (TEQSA). In fact, the promotion of English language standards (or academic literacy) is now seen as encompassing the whole student body, rather than just international students (see, e.g., Arkoudis et al. 2012).

From an assessment perspective, the two most relevant Good Practice Principles are these:

1. Universities are responsible for ensuring that their students are sufficiently competent in English to participate effectively in their university studies.
2. Students' English language development needs are diagnosed early in their studies and addressed, with ongoing opportunities for self-assessment (AUQA 2009, p. 4).

A third principle, which assigns shared responsibility to the students themselves, should also be noted:

3. Students have responsibilities for further developing their English language proficiency during their study at university and are advised of these responsibilities prior to enrolment. (ibid.)

These principles have produced initiatives in many Australian institutions to design what have become known as post-entry language assessments (PELAs). Actually, a small number of assessments of this kind pre-date the developments of the last 10 years, notably the Diagnostic English Language Assessment (DELA) at the University of Melbourne (Knoch et al. this volume) and Measuring the Academic Skills of University Students at the University of Sydney (Bonanno and Jones 2007). The more recent developments have been documented in national surveys by Dunworth (2009) and Dunworth et al. (2013). The latter project led to the creation of the Degrees of Proficiency website (www.degreesofproficiency.aall.org.au), which offers a range of useful resources on implementing the Good Practice Principles, including a database of PELAs in universities nationwide.

In New Zealand, although the term PELA is not used, the University of Auckland has implemented the most comprehensive assessment programme of this kind, the

Diagnostic English Language Needs Assessment (DELNA), which has been in operation since 2002 (see Read and von Randow this volume). Currently all first-year undergraduate students and all doctoral candidates are screened through DELNA, regardless of their language background. The impetus for the development of the programme came from widespread perceptions among staff of the university in the 1990s that many students were inadequately prepared for the language and literacy demands of their studies. Attention was focused not simply on international students but a range of other groups in the student population, including permanent residents who had immigrated relatively recently; mature students with no recent experience of formal study; and ethnic minority students admitted on equity grounds. Even mainstream students could no longer be assumed to have an acceptable level of academic literacy. There were legislative constraints on singling out particular groups of domestic students for English assessment, and so the University eventually required that all incoming students should be screened.

Dunworth's surveys in Australia have revealed that PELAs and the institutional policies associated with them take rather different forms from one institution to another. Nevertheless, it is possible to list a number of distinctive features that this kind of assessment may have, in the original context of Australia and New Zealand.

- Although international students are often the primary target group for assessment, some PELAs are administered more widely to incoming students with English as an additional language, whether they be international or domestic in origin. Given the diversity of language backgrounds and educational experiences among today's students, any division according to the old dichotomies of non-native vs. native or non-English- vs. English-speaking background may be quite problematic and seen rightly or wrongly as discriminatory.
- A related issue is whether it should be made mandatory for the targeted students to undertake the assessment, with sanctions for non-compliance – or whether the PELA should be made available to students with varying degrees of encouragement or persuasion to take it. There is often some resistance from academics to the idea of compulsion, on the basis that it is unreasonable to oblige students who have already met the university's matriculation requirements to take a further assessment.
- In any event the assessment is administered after students have been admitted to the university and, no matter how poorly they perform, they will not be excluded from degree study on the basis of the PELA result.
- A PELA is usually developed within the institution where it is used, although some universities have pooled their expertise and others have licensed the assessment from another university. It is funded by the university, or in some cases by a particular faculty, at no cost to the student.
- PELAs typically target reading, writing and listening skills, but often include measures of language knowledge (vocabulary or grammar items; integrated formats such as the cloze procedure), which are seen as adding diagnostic value to the assessment.
- The assessment should be closely linked to the opportunities available on campus for students to enhance their academic language proficiency or literacy,

through credit courses in ESL, academic English or academic writing; short courses and workshops; online study resources; tutoring services; peer mentoring; and so on. In some academic programmes, language support is embedded in the teaching of particular first-year subject courses.

- In some cases, students are required to take a credit course if their PELA results are low. Otherwise (or in addition), the assessment is seen as more diagnostic in nature, and the reporting of their results is accompanied by advice on how to improve their language skills.

This cluster of characteristics shows how a PELA is distinct from the major proficiency tests like IELTS and TOEFL, which govern the admission of international students to English-medium universities.

A PELA may be more similar to a placement test administered to students at a particular institution. However, many placement tests are designed simply to assign incoming students to a class at the appropriate level of an English language or writing/composition programme as efficiently as possible, which means that they lack the range of features – and the underlying philosophy – of a PELA, as outlined above. It is worth noting that two major recent survey volumes on language assessment (Fulcher and Davidson 2012; Kunnan 2014) barely mention placement tests at all, whereas the chapter by Green (2012) in a third volume states that "Ultimately, for a placement test to fulfill its purpose its use must result in a satisfactory assignment of learners to classes – at least in terms of language level" (p. 166). A PELA mostly has broader ambitions than this.

The phenomenon of post-entry language assessment is discussed in much greater depth in my book *Assessing English proficiency for university study* (Read 2015), including both the Australian and New Zealand experience and case studies of similar assessments in other countries as well. The present volume complements the earlier one by presenting a series of research studies on these kinds of assessment from universities in Australia, New Zealand, Canada, the United States, Hong Kong, Oman and South Africa. I have chosen to use the term "post-admission assessment" for the title to make the point that this volume is not primarily about the Australian experience with PELA but rather it ranges more widely across a variety of national contexts in which English-medium higher education occurs.

2 Structure of the Volume

2.1 Implementing and Monitoring Undergraduate Assessments

The first four chapters of the book, following this one, focus on the assessment of students entering degree programmes in English-medium universities at the *undergraduate* level. The **second chapter**, by **Ute Knoch**, **Cathie Elder and Sally O'Hagan**, discusses recent developments at the University of Melbourne, which has been a pioneering institution in Australia in the area of post-admission

assessment, not only because of the high percentage of students from non-English-speaking backgrounds on campus but also because the establishment of the Language Testing Research Centre (LTRC) there in 1990 made available to the University expertise in test design and development. The original PELA at Melbourne, the Diagnostic English Language Assessment (DELA), which goes back to the early 1990s, has always been administered on a limited scale for various reasons. A policy was introduced in 2009 that all incoming undergraduate students whose English scores fell below a certain threshold on IELTS (for international students) or the Victorian matriculation exams (domestic students) would be required to take DELA, followed by an academic literacy development programme as necessary (Ransom 2009). However, it has been difficult to achieve full compliance with the policy. This provided part of the motivation for the development of a new assessment, now called the Post-admission Assessment of Language (PAAL), which is the focus of the Knoch et al. chapter.

Although Knoch et al. report on a trial of PAAL in two faculties at Melbourne University, the assessment is intended for use on a university-wide basis and thus it measures general academic language ability. By contrast, in Chap. 3 **Janna Fox**, **John Haggerty and Natasha Artemeva** describe a programme tailored specifically for the Faculty of Engineering at Carleton University in Canada. The starting point was the introduction of generic screening measures and a writing task licensed from the DELNA programme at the University of Auckland in New Zealand (discussed in Chap. 7), but as the Carleton assessment has evolved, it was soon recognised that a more discipline-specific set of screening measures was required to meet the needs of the faculty. Thus, both the input material and the rating criteria for the writing task were adapted to reflect the expectations of engineering instructors, and recently a more appropriate reading task and a set of mathematical problems have been added to the test battery. Another feature of the Carleton programme has been the integration of the assessment with the follow-up pedagogical support. This has involved the embedding of the assessment battery into the delivery of a required first-year engineering course and the opening of a support centre staffed by upper-year students as peer mentors. Fox et al. report that there is a real sense in which students in the faculty have taken ownership of the centre, with the result that it is not stigmatised as a remedial place for at-risk students, but somewhere where a wide range of students can come to enhance their academic literacy in engineering.

The term academic literacy is used advisedly here to refer to the discipline-specific nature of the assessment at Carleton, which distinguishes it from the other programmes presented in this volume; otherwise, they all focus on the more generic construct of academic language proficiency (for an extended discussion of the two constructs, see Read 2015). The one major example of an academic literacy assessment in this sense is Measuring the Academic Skills of University Students (MASUS) (Bonanno and Jones 2007), a procedure developed in the early 1990s at the University of Sydney, Australia, which involves the administration of a discipline-specific integrated writing task. This model requires the active involvement of instructors in the discipline, and has been implemented most effectively in

professional fields such as accountancy, architecture and pharmacy. However, PELAs are generally designed for students entering degree programmes across the university and often target those who are linguistically at risk through their limited competence in the lexical, grammatical and discoursal systems of the language – hence the diagnostic function of the assessment tasks.

We next move beyond the traditional English-speaking countries. The fourth and fifth chapters focus on post-admission assessments in Hong Kong, now a Special Administrative Region of China but under British administration for more than a century until 1997. Thus, English has had a primary role in the public domain and the commercial life of Hong Kong, even though the population is predominantly Cantonese-speaking. Both before and after the transfer of sovereignty, the issue of the medium of instruction in Hong Kong schools has been a matter of ongoing debate and controversy (Evans 2002; So 1989). From 1997, mother tongue education in Cantonese was strongly promoted for most schools but, faced with ongoing demand from parents for English-medium education, in 2010 the Government discontinued the practice of classifying secondary schools as English-medium or Chinese-medium in favour of a "fine-tuning" policy which allowed schools to adopt English as a medium to varying degrees, depending on the students' ability to learn through English, the teachers' capability to teach through English and the availability of support measures at the school (Education Bureau 2009). However, the debate continues over the effectiveness of the new policy in raising the standard of English among Hong Kong secondary students (Chan 2014; Poon 2013).

This obviously has flow-on effects at the university level, as Hong Kong has moved to broaden participation in higher education beyond the elite group of students from schools with a long tradition of English-medium study. There are now eight public universities in the SAR, and all but one (the Chinese University of Hong Kong) are English-medium institutions. Responding to the concerns of employers about the English communication skills of graduating students, there has been a focus on finding an appropriate and acceptable exit test for those completing their degree studies (Berry and Lewkowicz 2000; Qian 2007), but clearly the situation also needs to be addressed on entry to the university as well. Another recent change has been the introduction of 4-year undergraduate degree programmes, rather than the traditional 3-year British-style degrees, and a consequent reduction in senior secondary schooling from 4 years to 3. The expectation is that this will increase the need for Hong Kong students to devote much of their first year of university study to enhancing their academic English skills.

This then is the background to the two chapters on Hong Kong. Chapter 4, by **Edward Li**, introduces the English Language Proficiency Assessment (ELPA), which has been developed for students entering the Hong Kong University of Science and Technology. Li emphasises the close link between the assessment and the coursework in academic English which the students undertake in their first year of study, as part of a major institutional commitment to improving English language standards. As he writes, "ELPA is curriculum-led and curriculum-embedded". Thus, it is a relatively comprehensive instrument, in its coverage of the four skills as well as vocabulary knowledge – comparable in scope to a communicative proficiency

test. Although it is not primarily designed as a diagnostic test battery, it is similar to other post-admission assessments in this volume in that, after the test results are released, students have an individual consultation with their class teacher to negotiate a plan for their English study for the remainder of the academic year. ELPA has also provided opportunities for teachers at the Center for Language Education to enhance their professional skills in assessment and to see for themselves the links between what is assessed and what they teach in the classroom.

Whereas most post-admission assessments are administered on a one-off basis at the beginning of the academic year, ELPA also functions as a higher-stakes achievement measure for HKUST students at the end of the first year. The other Hong Kong measure, the Diagnostic English Language Tracking Assessment (DELTA), is even more longitudinal in nature, in that it is intended to be taken by the students in each year of their undergraduate study, as a tool for them to monitor their progress in enhancing their language knowledge and receptive skills for study purposes. As **Alan Urmston**, **Michelle Raquel and Vahid Aryadoust** explain in Chap. 5, the Hong Kong Polytechnic University is the lead institution in implementing DELTA, along with three other partner universities in Hong Kong. Policies vary from one participating university to another on whether students are required to take DELTA and how the assessment is linked to the academic English study programmes which each institution provides. This means that the design of the instrument is not tied to a particular teaching curriculum, as ELPA is, but it draws on key components of the construct of language knowledge, as defined by Bachman and Palmer (2010), within an academic study context. DELTA is a computer-based assessment which makes sophisticated use of the Rasch Model to evaluate the quality of the test items and to provide a basis for interpreting the scores for the students and other stakeholders. This includes a DELTA Track, which takes account of the student's past performance and maps the trajectory to a new target level of achievement which the student can set. Thus, it is designed to facilitate individualised learning, to complement the students' formal course work in English.

2.2 Addressing the Needs of Doctoral Students

The second section of the book includes two studies of post-admission assessments for postgraduate students, and more specifically doctoral candidates. Although international students at the postgraduate level have long been required to achieve a minimum score on a recognised English proficiency test for admission purposes, normally this has involved setting a somewhat higher score on a test that is otherwise the same as for undergraduates. However, there is growing recognition of the importance of addressing the specific language needs of doctoral students, particularly in relation to speaking and writing skills. Such students have usually developed academic literacy in their discipline through their previous university education but, if they are entering a fully English-medium programme for the first time, their doctoral studies will place new demands on their proficiency in the language.

In major US universities with strong programmes in the sciences and engineering, International Teaching Assistants (ITA) have had a prominent position since at least the 1980s, taking instructional roles in undergraduate courses in return for financial support to pursue their own graduate studies. This means that they need good oral-aural ability as well as basic teaching skills. In fact, in numerous US states the legislature has mandated the assessment and training of prospective ITAs in response to public concerns about their competence to perform their teaching roles. Some universities use existing tests, such as the speaking section of the internet-based TOEFL (iBT) (Xi 2008), but others have developed their own in-house instruments. One well-documented case is the Oral English Proficiency Test (OEPT) at Purdue University in Indiana, which is the focus of Chap. 6 by **Xun Yan**, **Suthathip Ploy Thirakunkovit**, **Nancy L. Kauper and April Ginther**. The test is closely linked to the Oral English Proficiency Program (OEPP), which provides training in the necessary presentational and interactional skills for potential ITAs whose OEPT score indicates that they lack such skills.

The OEPT is the only post-admission assessment represented in this volume which tests oral language ability exclusively. Generally, speaking is assigned a lower priority than other skills, having regard for the time and expense required to assess oral skills reliably. However, an oral assessment is essential for prospective ITAs and the solution adopted for the OEPT, which is administered to 500 graduate students a year, was to design a computer-based semi-direct test in which the test-takers respond to prompts on screen rather than interacting with a human interlocutor. An important feature of the assessment for those students placed in the OEPP on the basis of a low test score is an individual conference with an instructor to review the student's performance on the test and set instructional goals for the semester. The close relationship between the assessment and the instruction is facilitated by the fact that OEPP instructors also serve as OEPT raters.

The OEPT is also distinct from the other assessments in that it assesses professional skills rather than ones required for academic study. Of course, the employment context for the ITAs is academic, and developing good oral skills will presumably stand them in good stead as graduate students, but they are primarily being assessed on their employability as instructors in the university, not their ability to cope with the language demands of their studies.

The following Chap. 7, by **John Read and Janet von Randow**, discusses a more general application of post-admission assessment to all incoming doctoral candidates at the University of Auckland in New Zealand. This involved not the development of a new instrument but the extension of the existing Diagnostic English Language Needs Assessment (DELNA), which was already well established for undergraduate students. DELNA is unusual among PELAs in Australian and New Zealand universities in that for several years it has been administered to all domestic and international first-year undergraduates, regardless of their language background. Since 2011, the same policy has been applied to all new doctoral students. The only innovation in the assessment for them has been an extended writing task.

As in the case of the OEPT at Purdue, the DELNA programme includes an individual advisory session for students whose assessment results show that they

are at risk of language-related difficulties in their studies. For doctoral candidates the advising process is more formal than for undergraduates and it results in the specification of language enrichment objectives which are reviewed at the end of each student's provisional year of registration, as part of the process to determine whether their candidacy should be confirmed. The DELNA requirement has been complemented with an augmented range of workshops, discussion groups, online resources and other activities tailored to the academic language needs of doctoral students. Interestingly, although speaking skills are not assessed in DELNA, Read and von Randow highlight the need for international doctoral candidates to develop their oral communication ability as a means to form what Leki (2007) calls "socio-academic relationships" through interaction with other students and with university staff. Both the literature and the interview data presented in Chap. 7 attest to the isolation that international PhD students can experience, especially when their studies do not involve any course work.

A third postgraduate assessment, the Test of Academic Literacy for Postgraduate Students (TALPS) in South Africa, is discussed later in Chap. 10.

2.3 Issues in Assessment Design

In the third section of the book, there are three chapters which shift the focus back to English-medium university education in societies where (as in Hong Kong) most if not all of the domestic student population are primary speakers of other languages. These chapters are also distinctive in the attention they pay to design issues in post-admission assessments.

Chapter 8, by **Thomas Roche**, **Michael Harrington**, **Yogesh Sinha and Christopher Denman**, investigates the use of a particular test format for the purposes of post-admission assessment at two English-medium universities in Oman, a Gulf state which came under British influence in the twentieth century but where English remains a foreign language for most of the population. The instrument is what the authors call the Timed Yes/No (TYN) vocabulary test, which measures the accuracy and speed with which candidates report whether they know each of a set of target words or not. Such a measure would not normally be acceptable in a contemporary high-stakes proficiency test, but it has a place in post-admission assessments. Vocabulary sections are included in the DELTA, DELNA and ELPA test batteries, and the same applies to TALL and TALPS (Chaps. 9 and 10). Well-constructed vocabulary tests are highly reliable and efficient measures which have been repeatedly shown to be good predictors of reading comprehension ability and indeed of general language proficiency (Alderson 2005). They fit well with the diagnostic purpose of many post-admission assessments, as distinct from the more communicative language use tasks favoured in proficiency test design. Roche et al. argue that a TYN test should be seriously considered as a cost-effective alternative to the existing resource-intensive placement tests used as part of the admission process to foundation studies programmes at the two institutions.

The TYN test trial at the two institutions produced promising results but also some reasons for caution in implementing the test for operational purposes. The vocabulary test was novel to the students not only in its Yes/No format but also the fact that it was computer-based. A comparison of student performance at the two universities showed evidence of a digital divide between students at the metropolitan institution and those at the regional one; there were also indications of a gender gap in favour of female students at the regional university. This points to the need to ensure that the reliability of placement tests and other post-admission assessments is not threatened by the students' lack of familiarity with the format and the mode of testing. It also highlights the value of obtaining feedback from test-takers themselves, as several of the projects described in earlier chapters have done.

The other two chapters in the section come from a team of assessment specialists affiliated to the Inter-institutional Centre for Language Development and Assessment (ICELDA), which – like the DELTA project in Hong Kong – involves collaboration among four participating universities in South Africa to address issues of academic literacy faced by students entering each of the institutions. The work of ICELDA is informed by the multilingual nature of South African society, as well as the ongoing legacy of the political and educational inequities inflicted by apartheid on the majority population of the country. This makes it essential that students who will struggle to meet the language demands of university study through the media of instruction of English or Afrikaans should be identified on entry to the institution and should be directed to an appropriate programme of academic literacy development.

Two tests developed for this purpose, the Test of Academic Literacy Levels (TALL) and its Afrikaans counterpart, the Toets van Akademiese Geletterdheidsvlakke (TAG), are discussed in Chap. 9 by **Albert Weideman**, **Rebecca Patterson and Anna Pot**. These tests are unusual among post-admission assessments in the extent to which an explicit definition of academic literacy has informed their design. It should be noted here that the construct was defined generically in this case, rather than in the discipline-specific manner adopted by Read (2015) and referred to in Chap. 3 in relation to the Carleton University assessment for engineering students. The original construct definition draws on current thinking in the applied linguistic literature, particularly work on the nature of academic discourse. The authors acknowledge that compromises had to be made in translating the components of the construct into an operational test design, particularly given the need to employ objectively-scored test items for practical reasons in such large-scale tests. The practical constraints precluded any direct assessment of writing ability, which many would consider an indispensable element of academic literacy.

In keeping with contemporary thinking about the need to re-validate tests periodically, Weideman et al. report on their recent exercise to revisit the construct, leading to some proposed new item types targeting additional components of academic literacy. One interesting direction, following the logic of two of the additions, is towards the production of some field-specific tests based on the same broad construct. It would be useful to explore further the diagnostic potential of these tests through the reporting of scores for individual sections, rather than just the total score. To date this potential has not been realised, largely again on the practical

grounds that more than 30,000 students need to be assessed annually, and thus overall cut scores are simply used to determine which lower-performing students will be required to take a 1-year credit course in academic language development.

This brings us to Chap. 10, by **Avasha Rambiritch and Albert Weideman**, which complements Chap. 9 by giving an account of the other major ICELDA instrument, the Test of Academic Literacy for Postgraduate Students (TALPS). As the authors explain, the development of the test grew out of a recognition that postgraduate students were entering the partner institutions with inadequate skills in academic writing. The construct definition draws on the one for TALL and TAG but with some modification, notably the inclusion of an argumentative writing task. The test designers decided that a direct writing task was indispensable if the test was to be acceptable (or in traditional terminology, to have face validity) to postgraduate supervisors in particular.

The last point is an illustration of the authors' emphasis on the need for test developers to be both transparent and accountable in their dealings with stakeholders, including of course the test-takers. At a basic level, it means making information easily available about the design of the test, its structure and formats, and the meaning of test scores, as well as providing sample forms of the test for prospective candidates to access. Although this may seem standard practice in high-stakes testing programmes internationally, Rambiritch and Weideman point out that such openness is not common in South Africa. In terms of accountability, the test developers identify themselves and provide contact details on the ICELDA website. They are also active participants in public debate about the assessment and related issues through the news media and in talks, seminars and conferences. Their larger purpose is to promote the test not as a tool for selection or exclusion but as one means of giving access to postgraduate study for students from disadvantaged backgrounds.

Although universities in other countries may not be faced with the extreme inequalities that persist in South African society, this concern about equity of access can be seen as part of the more general rationale for post-admission language assessment and the subsequent provision of an "intervention" (as Rambiritch and Weideman call it), in the form of opportunities for academic language development. The adoption of such a programme signals that the university accepts a responsibility for ensuring that students it has admitted to a degree programme are made aware of the fact that they may be at risk of underachievement, if not outright failure, as a result of their low level of academic language proficiency, even if they have met the standard requirements for matriculation. The institutional responsibility also extends to the provision of opportunities for students to enhance their language skills, whether it be through a compulsory course, workshops, tutorial support, online resources or peer mentoring.

The concluding Chap. 11, by **John Read**, discusses what is involved for a particular university in deciding whether to introduce a post-admission language assessment, as part of a more general programme to enhance the academic language development of incoming students from diverse language backgrounds. There are pros and cons to be considered, such as how the programme will be viewed

externally and whether the benefits will outweigh the costs. Universities are paying increasing attention to the employability of their graduates, whose attributes are often claimed to include effective communication ability. This indicates that both academic literacy and professional communication skills need to be developed not just in the first year of study but throughout students' degree programmes. Thus, numerous authors now argue that language and literacy enhancement should be embedded in the curriculum for all students, but there are daunting challenges in fostering successful and sustained collaboration between English language specialists and subject teaching staff. The chapter concludes by exploring the ideas associated with English as a Lingua Franca (ELF) and how they might have an impact on the post-admission assessment of students.

3 Broad Themes in the Volume

To conclude this introduction, I will identify three themes which each go across several chapters in the volume.

3.1 Validation of Post-Admission Assessments

As with any assessment, a key question with PELAs is how to validate them. The authors of this volume have used a variety of frameworks and conceptual tools for this purpose, especially ones which emphasise the importance of the consequences of the assessment. This is obviously relevant to post-admission assessment programmes, where by definition the primary concern is not only to identify incoming students with academic language needs but also to ensure that subsequently they have the opportunity to develop their language proficiency in ways which will enhance their academic performance at the university.

In Chap. 2, Knoch, Elder and O'Hagan present a framework which is specifically tailored for the validation of post-admission assessments. The framework is an adapted version of the influential one in language testing developed by Bachman and Palmer (2010), which in turn draws on the seminal work on test validation of Samuel Messick and more particularly the argument-based approach advocated by Michael Kane. It specifies the sequence of steps in the development of an argument to justify the interpretation of test scores for a designated purpose, together with the kinds of evidence required at each step in the process. The classic illustration of this approach is the validity argument for the internet-based TOEFL articulated by Chapelle et al. (2008). Knoch and Elder have applied their version of the framework to several PELAs and here use it as the basis for evaluating the Post-entry Assessment of Academic Language (PAAL) at the University of Melbourne.

An alternative approach to validation is Cyril Weir's (2005) socio-cognitive model, which incorporates the same basic components as the Bachman and Palmer

framework, including an emphasis on consequential validity as part of a consideration of the social context in which the assessment occurs. Weir's model is promoted as being a more practical tool for operational use in test development projects than the more academically-oriented Bachman and Palmer framework. It has been prominent in British language testing, particularly for validating the main suite examinations of Cambridge English Language Assessment. While acknowledging the influence of Messick's ideas, Li has used the Weir model in Chap. 4 as the basis for identifying relevant evidence for the consequential validity of the English Language Proficiency Assessment (ELPA) at the Hong Kong University of Science and Technology.

At Purdue University the Oral English Proficiency Program (as described by Yan et al. in Chap. 6) has adopted a quality management process to facilitate a periodic review of the quality of its assessment procedures. Here again Cambridge English Language Assessment have been leaders internationally in introducing quality control principles into the evaluation of their examination programmes and, as Saville (2012) explains, it complements the validation process by focusing on the operational areas of test production and administration. The OEPP case illustrates in particular how feedback from test-takers can serve the ongoing process of quality control.

A fourth framework, which is not specific to language assessment, has been employed in Chap. 3 by Fox, Haggerty and Artemeva. They evaluated the diagnostic assessment procedure for engineering students by means of John Creswell's multistage evaluation design. Not surprisingly, given Creswell's long-term promotion of mixed-methods research, this involved gathering both quantitative and qualitative evidence through a longitudinal series of projects over a 6-year period. Technically, this can be seen as a programme evaluation rather than simply a validation of the assessment, but the fact that the assessment is embedded so fully in first-year engineering studies in the faculty means that it is an appropriate approach to take.

Finally, the other main viewpoint on test validity is provided by Weideman, Patterson and Pot in their work on the construct definition of academic literacy, which underlies the design of the Test of Academic Language Levels (TALL) and its companion instruments (Chap. 9).

3.2 Feedback from Test-Takers

A notable feature of several studies in the volume is the elicitation of feedback from students who have taken the assessment. This can be seen as a form of validity evidence or, as we have just seen in the case of the OEPP at Purdue University, as input to a quality management procedure. Although an argument can be made for obtaining test-taker views at least at the development stage of any language testing programme, it is particularly desirable for a post-admission assessment for three reasons. First, like a placement test, a PELA is administered shortly after students

arrive on campus and, if they are uninformed or confused about the nature and purpose of the assessment, it is less likely to give a reliable measure of their academic language ability. The second reason is that the assessment is intended to alert students to difficulties they may face in meeting the language demands of their studies and often to provide them with beneficial diagnostic feedback. This means that taking the assessment should ideally be a positive experience for them and anything which frustrates them about the way the test is administered or the results are reported will not serve the intended purpose. The other, related point is that the assessment is not an end in itself but should be the catalyst for actions taken by the students to enhance their academic language ability. Thus, feedback from the students after a period of study provides evidence as to whether the consequences of the assessment are positive or not, in terms of what follow-up activities they engage in and what factors may inhibit their uptake of language development opportunities.

Feedback from students was obtained in different ways and for various purposes in these studies. In the development of the PAAL (Chap. 2), a questionnaire was administered shortly after the two trials, followed later by focus groups. A similar pattern of data-gathering was conducted in the studies of the two Hong Kong tests, ELPA (Chap. 4) and DELTA (Chap. 5). On the other hand, the Post Test Questionnaire is incorporated as a routine component of every administration of the OEPT (Chap. 6), to monitor student reactions to the assessment on an ongoing basis. A third model is implemented for DELNA (Chap. 7). In this case, students are invited to complete a questionnaire and then participate in an interview only after they have completed a semester of study. Although this voluntary approach greatly reduces the response rate, it provides data on the students' experiences of engaging (or not) in academic language enhancement activities as well as their reactions to the assessment itself.

3.3 The Diagnostic Function

One characteristic of post-admission assessments which has already been referred to is their potential for providing diagnostic information about the students' academic language ability. In fact, three of the assessments (DELA, DELNA and DELTA) include the term "diagnostic" in their names. Diagnosis has become a hot topic in language assessment in the last 10 years, stimulated by the work of Charles Alderson and his colleagues (2014, 2015) and by Alderson (2005), as well as other scholars (see e.g. Lee 2015). There is continuing debate about the nature of diagnosis and how it should be carried out in language teaching and assessment. A test administered to thousands of students is on the face of it diagnostic in a different sense than a procedure conducted one-on-one in a classroom setting by a teacher with special skills.

In Chap. 3, Fox et al. stress the importance of the connection between the assessment and pedagogical intervention. They argue that "a diagnostic assessment proce-

dure cannot be truly diagnostic unless it is linked to feedback, intervention, and support" (p. xx). For them, this entails the production of a learning profile which leads to individually tailored language support for the engineering students. Similarly, for Knoch et al. (Chap. 2), a key diagnostic element is a subskill profile, which goes beyond a single overall score and provides the basis for subsequent student advising. They report that student participants in their trials of PAAL complained that the reported results lacked detail about their strengths and weaknesses.

The analytic rating scales that are typically used for writing and speaking assessment lend themselves well to diagnostic reporting and advising, as noted by Knoch et al. and by Yan et al. in Chap. 6, with regard to the Oral English Proficiency Test (OEPT). Although the OEPT score is reported on a single scale, the class teachers in the OEPP have access to the analytic ratings, which they review in individual conferences with students in their class. In the case of ELPA at Hong Kong University of Science and Technology (Chap. 4), Li writes that the assessment was not designed to be diagnostic, but the results are reported as criterion-referenced levels of performance in the skill areas using can-do statements and, as happens with the assessments which are more explicitly diagnostic in nature, the performance descriptors are used in consultations between class teachers and students about English learning plans for the first year of study. On the other hand, in Chap. 10 Rambiritch and Weideman explain that the Test of Academic Literacy for Postgraduate Students has diagnostic potential, in that its clusters of multiple-choice items each assess one component of the authors' academic literacy construct; however, the potential is currently not realised in practice because of resource constraints. The results are reported simply on a scale of the level of risk the student faces as a result of inadequate literacy.

The most comprehensive statement of the diagnostic nature of the assessment is found in Chap. 5, where Urmston et al. explain how DELTA adheres to five tentative principles of language diagnosis articulated by Alderson et al. (2014, 2015). These include: providing an interactive report for users to interpret for their own purposes; offering a user-friendly computer interface; incorporating diverse stakeholder views, and student self-assessment in particular; embedding the assessment in a whole diagnostic process which leads to decisions negotiated between student and lecturer; and making a strong link between the assessment and appropriate learning activities in the future.

Overall, the various conceptions of diagnosis serve to highlight the fundamental point that a post-admission language assessment is not an end in itself but a means of encouraging, if not requiring, students at risk of poor academic performance to enhance their academic language proficiency through the various resources available to them.

References

Alderson, J. C. (2005). *Diagnosing foreign language proficiency*. London: Continuum.

Alderson, J. C., Brunfaut, T., & Harding, L. (2015). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. *Applied Linguistics, 36*, 236–260.

Alderson, J. C., Haapakangas, E.-L., Huhta, A., Nieminen, L., & Ullakonoja, R. (2014). *The diagnosis of reading in a second or foreign language*. New York: Routledge.

Arkoudis, S., Baik, C., & Richardson, S. (2012). *English language standards for higher education: From entry to exit*. Camberwell: ACER Press.

Australian Education International (AEI). (2007). *Final report. Outcomes from a national symposium: English language competence of international students*. Retrieved April 20, 2011, from: http://aei.gov.au/AEI/PublicationsAndResearch/Publications/NS_Outcomes_Syposium_pdf.pdf

Australian Universities Quality Agency (AUQA). (2009, March). *Good practice principles for English language proficiency for international students in Australian Universities*. Retrieved 21 Jan 2014 from: http://www.aall.org.au/sites/default/files/Final_Report-Good_Practice_Principles2009.pdf.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Berry, V., & Lewkowicz, J. (2000). Exit tests: Is there an alternative? *Hong Kong Journal of Applied Linguistics, 5*(1), 19–49.

Birrell, B. (2006). Implications of low English standards among overseas students at Australian universities. *People and Place, 14*(4), 53–64. Melbourne: Centre for Population and Urban Research, Monash University.

Bonanno, H., & Jones, J. (2007). *The MASUS procedure: Measuring the academic skills of university students. A resource document*. Sydney: Learning Centre, University of Sydney. http://sydney.edu.au/stuserv/documents/learning_centre/MASUS.pdf

Chan, J. Y. H. (2014). Fine-tuning language policy in Hong Kong education: Stakeholders' perceptions, practices and challenges. *Language and Education, 28*(5), 459–476.

Chapelle, C. A., Enright, M., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the test of english as a foreign language*. New York: Routledge.

Crystal, D. (2003). *English as a global language* (2nd ed.). Cambridge: Cambridge University Press.

Dunworth, K. (2009). An investigation into post-entry English language assessment in Australian universities. *Journal of Academic Language and Learning, 3*(1), 1–13.

Dunworth, K. Drury, H., Kralik, C., Moore, T., & Mulligan, D. (2013). *Degrees of proficiency: Building a strategic approach to university students' English language assessment and development*. Sydney: Australian Government Office for Learning and Teaching. Retrieved April 15, 2014, from www.olt.gov.au/project-degrees-proficiency-building-strategic-approach-university-studentsapos-english-language-ass

Education Bureau. (2009). *Fine-tuning the medium of instruction for secondary schools* (Education Bureau Circular No. 6/2009). Hong Kong: Government of the Hong Kong Special Administrative Region. Retrieved April 16, 2015, from http://www.edb.gov.hk/en/edu-system/primary-secondary/applicable-to-secondary/moi/

Evans, S. (2002). The medium of instruction in Hong Kong: Policy and practice in the new English and Chinese streams. *Research Papers in Education, 17*(1), 97–120.

Fulcher, G., & Davidson, F. (Eds.). (2012). *The Routledge handbook of language testing*. London: Routledge.

Green, A. (2012). Placement testing. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge guide to second language assessment* (pp. 164–170). New York: Cambridge University Press.

Jenkins, J. (2007). *English as a lingua franca: Attitudes and identity*. Oxford: Oxford University Press.

Kunnan, A. J. (Ed.). (2014). *The companion to language assessment*. Chichester: Wiley Blackwell.

Lee, Y. W. (Ed.) (2015). Special issue: Future of diagnostic language testing. *Language Testing, 32*(3), 293–418.

Leki, I. (2007). *Undergraduates in a second language: Challenges and complexities of academic literacy development*. Mahwah: Lawrence Erlbaum.

Murray, N. (2016). *Standards of English in higher education: Issues, challenges and strategies*. Cambridge: Cambridge University Press.

Phillipson, R. (2009). English in higher education: Panacea or pandemic? In R. Phillipson (Ed.), *Linguistic imperialism continued* (pp. 195–236). New York: Routledge.

Poon, A. Y. K. (2013). Will the new fine-tuning medium-of-instruction policy alleviate the threats of dominance of English-medium instruction in Hong Kong? *Current Issues in Language Planning, 14*(1), 34–51.

Qian, D. D. (2007). Assessing university students: Searching for an English language exit test. *RELC Journal, 38*(1), 18–37.

Ransom, L. (2009). Implementing the post-entry English language assessment policy at the University of Melbourne: Rationale, processes and outcomes. *Journal of Academic Language and Learning, 3*(2), 13–25.

Read, J. (2015). *Assessing English proficiency for university study*. Basingstoke: Palgrave Macmillan.

So, D. W. C. (1989). Implementing mother-tongue education amidst societal transition from diglossia to triglossia in Hong Kong. *Language and Education, 3*(1), 29–44.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

Xi, X. (2008). *Investigating the criterion-related validity of the TOEFL speaking scores for ITA screening and setting standards for ITAs* (ETS research reports, RR-08-02). Princeton: Educational Testing Service.

Part II
Implementing and Monitoring Undergraduate Assessments

Chapter 2
Examining the Validity of a Post-Entry Screening Tool Embedded in a Specific Policy Context

Ute Knoch, Cathie Elder, and Sally O'Hagan

Abstract Post-entry English language assessments (PELAs) have been instituted in many higher education contexts for the purpose of identifying the language needs of the linguistically and culturally diverse population of students entering English-medium universities around the world. The current chapter evaluates the validity of the Post-entry Assessment of Academic Language (PAAL), a PELA screening test trialled at two faculties (Engineering and Commerce) at a large Australian university. The chapter follows the approach to building an interpretative validity argument as outlined by Knoch and Elder (2013) and, by way of validity evidence, draws on data from the development phase of the test, its trial implementation and the evaluation phase of the trial. Evidence gathered during the development phase supports the first three inferences in the validity argument (evaluation, generalizability and explanation) and shows that the PAAL is technically adequate, relevant to the academic domain and effective in balancing the demands of efficiency and diagnostic sensitivity. Data from the trial however reveal problems in relation to the final two inferences in the validity argument concerning the relevance and appropriateness of decisions based on test scores and the consequences or perceived consequences of test use for the stakeholders involved. It is shown that these problems stem mainly from limitations in the institutional policy. We conclude that implementing the same test in a more hospitable policy setting might produce very different outcomes and assert the importance of evaluating tests in their policy contexts.

Keywords Post-entry language assessment • Screening test • Test validation • PAAL • English for academic purposes • Higher education

U. Knoch (✉) • C. Elder • S. O'Hagan
Language Testing Research Centre, University of Melbourne, Melbourne, Australia
e-mail: uknoch@unimelb.edu.au; caelder@unimelb.edu.au; sohagan@unimelb.edu.au

J. Read (ed.), *Post-admission Language Assessment of University Students*,
English Language Education 6, DOI 10.1007/978-3-319-39192-2_2

1 Introduction

While universities in major English-speaking countries have long-established English language entry requirements for international students in the form of a minimum score on an admissions test such IELTS or TOEFL, there is growing concern amongst academics that these minimum cut-scores may be too low to ensure that incoming students can cope with the language demands of their university study (Ginther and Elder 2014). Raising entry thresholds may not be deemed acceptable however, given the university's reliance on revenue from these fee-paying students on the one hand and, on the other, the fear of excluding otherwise academically talented students whose language skills might be expected to improve over time. In any case, as access to higher education broadens and diversifies, it has become evident that increasing numbers of domestic students, whether from English or non-English speaking backgrounds, may also be linguistically at risk in their academic study – even more so in some cases than their international student counterparts (Read 2015). Since limited command of English is associated with low retention rates and poor academic outcomes, addressing the issue has become a priority. In response to this situation many universities in Australia and elsewhere have instituted a Post-entry English Language Assessment (PELA) to identify the language needs of those who have been admitted to the university so that appropriate strategies can be devised to enhance their chances of academic success (Dunworth 2009).

Given the diverse range of students who may experience difficulties with academic English and the variable nature of the difficulties they face, the validity challenges for PELA design and use are considerable. A PELA needs to

- target the full range of potentially at risk students so that the chances of individuals falling through the net are minimized;
- capture relevant information from these individuals so that their particular language needs are clear;
- communicate this information to all relevant parties in a meaningful, timely and sensitive fashion;
- provide opportunities for the identified needs to be addressed; and, ultimately,
- consider whether the testing initiative is fulfilling its intended aim of improving student outcomes.

Different PELA models have been adopted by different institutions but evidence for their validity and efficacy is scarce and has been gathered in a piecemeal fashion. A paper by Knoch and Elder (2013) outlines a framework for PELA validation activity drawing on the influentialargument-based models proposed by Kane (1992, 2006) and Bachman and Palmer (2010). The framework sets out a series of inferences for which supporting evidence needs to be collected to claim that the scores and score interpretations based on a PELA are valid. For this, Knoch and Elder (2013) formulated a series of warrants for each inference in the PELA interpretive argument (evaluation, generalizability, explanation and extrapolation, decisions and consequences) which might be applicable to a range of different PELA contexts.

The authors also showed how the institutional policy determines how the decisions and consequences associated with a certain PELA play out. This crucial role of policy in PELA implementation will be taken up later in this chapter.

This framework has subsequently been applied to the evaluation of the Diagnostic English Language Needs Assessment at the University of Auckland (Read 2015) and also to two well-established Australian PELAs claiming either explicitly or implicitly to be diagnostic in their orientation: the Measuring the Academic Skills of University Students (MASUS) test at the University of Sydney (Bonanno and Jones 2007) and the Diagnostic English Language Assessment (DELA) at the University of Melbourne (Knoch and Elder 2016). In the latter study, Knoch and Elder consider the different inferences (evaluation, generalization, explanation/extrapolation, decisions and consequences) that underpin claims about diagnostic score interpretation in a PELA context and the associated warrants for which evidential support is required. The findings of the evaluation revealed that support for some of these warrants is lacking and that neither instrument can claim to be fully diagnostic. Although each PELA was found to have particular strengths, the claim that each provides diagnostic information about students which can be used as a basis for attending to their specific language needs is weakened by particular features of the assessment instruments themselves, or by the institutional policies determining the manner in which they are used. The rich diagnostic potential of the MASUS was seen to be undermined by limited evidence for reliability and also by the lack of standardized procedures for administration. The DELA, while statistically robust and potentially offering valid and useful information about reading, listening and writing sub-skills, was undermined by the policy of basing support recommendations on the student's overall score rather than on the sub-skill profile. The authors concluded that the DELA 'really functions as a screening test to group students into three broad bands – at risk, borderline or proficient – and there is no obvious link between the type of support offered and the particular needs of the student' (p. 16).

A further problem with both of these PELAs is that they are not universally applied: MASUS is administered only in certain departments and DELA is administered only to categories of students perceived to be academically at risk. Furthermore, there are few or no sanctions imposed on students who fail to sit the test, on the one hand, or fail to follow the support recommendations, on the other.

Similar problems of uptake were identified with the Diagnostic English Language Needs Assessment (DELNA) at the University of Auckland, a two-tiered instrument which includes both an initial screening test involving indirect, objectively-scored items and a follow-up performance-based diagnostic component, with the latter administered only to those falling below a specified threshold on the former. One of the challenges faced was convincing students who had performed poorly on the screening component to return for their subsequent diagnostic assessment (Elder and von Randow 2008). While uptake of the diagnostic component has improved over time, as awareness of the value of the DELNA initiative has grown, the success of the two-tiered system has been largely to due to the institution providing resources for a full-time manager, a half-time administrator and a half-time adviser whose

jobs include raising awareness of the initiative among academic staff and students, pursuing those who failed to return for the second round of testing, and offering one on one counseling on their English needs as required.

Given that many institutions are unwilling to make a similar commitment of resources to any PELA initiative, an alternative approach, which attempts to build on the strengths of previous models and to address their weaknesses, was devised by the authors of this chapter.

This paper offers an evaluation of the resulting instrument, known as the Post-entry Assessment of Academic Language (PAAL), based on the PELA evaluation framework referred to above. PAAL is the name adopted for a form of an academic English screening test known historically, and currently in other contexts, as the Academic English Screening Test (AEST). PAAL is designed to provide a quick and efficient means of identifying those students in a large and linguistically diverse student population who are likely to experience difficulties coping with the English language demands of academic study, while at the same time providing some diagnostic information. In the interests of efficiency it combines features of the indirect screening approach adopted at the University of Auckland (Elder and von Randow 2008) with a single task designed to provide some, albeit limited, diagnostic information, removing the need for a second round of assessment. It is based on the principle of universal testing, to allow for all at risk students to be identified, rather than targeting particular categories of students (as was the case for DELA) and builds on over 10 years of development and research done at the University of Melbourne and the University of Auckland.

2 Background to the Development and Format of the AEST/ PAAL

The AEST/PAAL was developed at the Language Testing Research Centre (LTRC) at the University of Melbourne in early 2009. It was initially commissioned for use by the University of South Australia; however, the rights to the test remain with the LTRC.

The test is made up of three sections: a text completion task, a speed reading task and an academic writing task, all completed within a 1-h time frame. (For a detailed account of the initial development and trialling, refer to Elder and Knoch 2009). The writing task was drawn from the previously validated DELA and the other two tasks were newly developed. The 20-min text completion task consists of three short texts and uses a C-test format (Klein-Braley 1985), in which every second word has been partially deleted. Students are required to reconstruct the text by filling in the gaps. The speed reading task, an adaptation of the cloze-elide format used for the screening component of DELNA at the University of Auckland (Elder and von Randow 2008), requires students to read a text of approximately 1,000 words in 10 min and identify superfluous words that have been randomly inserted. The writing task is an argumentative essay for which students are provided with a topic and have 30 min

to respond with 250–300 words (see Elder et al. 2009 for further detail). The text completion and speed reading tasks, used for screening purposes, are objectively scored and the writing task, intended for diagnostic use, is scored by trained raters using a three-category analytic rating scale. The test scores from the two screening components place students in one of three support categories as follows:

- Proficient. Students scoring in the highest range are deemed to have sufficient academic English proficiency for the demands of tertiary study.
- Borderline. Students scoring in the middle range are likely to be in need of further language support or development.
- At Risk. Students scoring in the lowest range are deemed likely to be at risk of academic failure if they do not undertake further language support and development.

It is recommended, for the sake of efficiency, that the writing component of the AEST/PAAL be completed by all students but be marked only for those scoring in the Borderline and At Risk categories on the two screening components. The writing thus serves to verify the results of the screening components for Borderline students (where categorization error is more likely) and potentially yields diagnostic information for less able students so that they can attend to their language weaknesses. The AEST/PAAL was however designed primarily as a screening test which efficiently groups students into the three categories above. For reasons of practicality and due to financial constraints the test was not designed to provide detailed feedback to test takers beyond the classification they are placed into and information about the support available to them on campus.

Following the development and trial of the prototype outlined in Elder and Knoch (2009), three more parallel versions of the test were developed (Knoch 2010a, b, 2011), initially for use at the University of South Australia, as noted above. In 2012, following a feasibility study (Knoch et al. 2012a), the University of Melbourne's English Language Development Advisory Group (ELDAG) supported a proposal by the LTRC to put the test online for eventual use at the University of Melbourne. This was funded in 2012 by a university Learning and Teaching Initiative Grant. The online platform was then developed by Learning Environments, a group of IT specialists supporting online learning and teaching initiatives at the University. Following the completion of the platform, the online delivery was tested on 50 test takers who had previously taken the University's DELA (Knoch et al. 2012b). The students also completed a questionnaire designed to elicit their experiences with the online platform. This small trial served as an extra check to verify cut-scores (between the Proficient, Borderline and At risk levels) which had been set during the development of the test using performance on the DELA as the benchmark (see below, and for further detail, Elder and Knoch 2009).

Based on the results of the small trial, a number of technical changes were made to the delivery of the test and in Semester 2, 2013, a trial on two full cohorts was undertaken (again funded by a Teaching and Learning Initiative grant) (Knoch and O'Hagan 2014). The trial targeted all newly incoming Bachelor of Commerce and Master of Engineering students as these two cohorts were considered to be

representative of students in very different disciplines and at the undergraduate and graduate levels. Following the trial, students were asked to complete an online questionnaire and a subset of students from both faculties took part in focus groups. For the purpose of the trial, the AEST was renamed as the Post-entry Assessment of Academic Language (PAAL) and this is the name we will use for the remainder of the chapter.

3 Methodology

As the overview of the historical development of the PAAL above shows, a number of trials and data collections have been conducted over the years. In this section, we will describe the following sources of data which we will draw on for this paper:

1. the PAAL development trial
2. the small trial of the online platform
3. the full trial on two student cohorts

3.1 PAAL Development Trial

There were 156 students who took part in the development trial of the PAAL, 71 from the University of South Australia and 85 from the University of Melbourne. All students were first year undergraduates and were from a range of L1 backgrounds, including a quarter from English-speaking backgrounds.

Test takers at both universities took the following components of the PAAL:

- Part A: Text completion (C-test with 4[1] texts of 25 items each) – 20 min
- Part B: Speed reading (Cloze elide with 75 items) – 10 min
- Part C: Writing task – 30 min

Test takers at the University of Melbourne had previously taken the DELA as well, and therefore recent scores for the following skills were also available:

- Reading: 46 item reading test based on two written texts – 45 min
- Listening: 30 item listening test based on lecture input – 30 min

3.2 Small Trial of the Online Platform

Fifty students from the University of Melbourne were recruited to take part in the small trial of the online platform developed by Learning Environments. The students completed a full online version of the PAAL from a computer or mobile

[1] A C-test with four texts is used for trial administrations whilst the final test form only includes three texts.

device at a place and time convenient to them. The final format of the PAAL adopted for the online trial and the full trial was as follows:

- Part A: Text completion (C-test with 3 texts of 25 items each) – 15 min
- Part B: Speed reading (Cloze elide with 75 items) – 10 min
- Part C: Writing task – 30 min

Following the trial, 49 of the participants completed an online questionnaire designed to elicit information about what device and browser they used to access the test, any technical issues they encountered, and whether the instructions to the test were clear and the timer visible at all times.

3.3 Full Trial on Two Student Cohorts

The full trial was conducted on two complete cohorts of commencing students at the beginning of Semester 2, 2013: Bachelor of Commerce (BCom) and Master of Engineering (MEng).

In the lead-up to the pilot implementation, extensive meetings were held with key stakeholders, in particular Student Centre staff from the respective faculties. It became evident very early in these discussions that universal testing is not possible as there is no mechanism to enforce this requirement on the students. Although it was not compulsory, all students in participating cohorts were strongly encouraged through Student Centre communications and orientation literature to complete the PAAL. At intervals during the pilot period, up to three reminder emails were sent by the respective Student Centres to remind students they were expected to take the test.

On completing the PAAL, all students were sent a report containing brief feedback on their performance and a recommendation for language support according to their results. The support recommendations were drafted in consultation with the Student Centres and Academic Skills[2] to ensure recommendations were in accord with appropriate and available offerings. The reports were emailed by the Language Testing Research Centre (LTRC) to each student within 1–2 days of their completing the PAAL. Cumulative spreadsheets of all students' results were sent by the LTRC to the Student Centres on a weekly basis throughout the pilot testing period.

The PAAL was taken by 110 BCom students, or 35 % of the incoming cohort of 310 students. In the MEng cohort, PAAL was taken by 60 students, comprising 12 % of the total of 491. The level of uptake for the BCom cohort was reported by the Commerce Student Centre as favourable compared with previous Semester 2 administrations of the DELA. Lower uptake for the MEng cohort was to be expected since traditionally post-entry language screening has not been required for graduate students at the University of Melbourne.

[2] The unit responsible for allocation and delivery of academic language support at the University.

The full trial was followed up with an evaluation in which feedback was sought from University stakeholders in order to develop recommendations for the best future form of the PAAL. Feedback came from students in the trial by means of a participant questionnaire and focus groups. Face-to-face and/or email consultation was used to gather feedback from Student Centres and Academic Skills.

Student consultation commenced with an online survey distributed to all pilot participants 2 weeks after taking the PAAL. Responses were received from 46 students, representing a 27 % response rate. Survey respondents were asked for feedback on the following topics: the information they received about the PAAL prior to taking the assessment; their experience of taking the PAAL (technical aspects, task instructions, face validity and difficulty of tasks); the PAAL report (results and recommendations); and the options for support/development and follow-up advice after taking the PAAL.

To gather more detailed feedback on these aspects of the PAAL, and to give students an opportunity to raise any further issues, four focus groups of up to 60 min duration were held: 20 students attended 1 of 4 faculty-specific focus groups, with an average of 5 students in each group. Group discussion was structured around the themes covered in the survey and participants were given the opportunity to elaborate their views and to raise any other issues relating to the PAAL that were of interest or concern to them.

4 Results

The remainder of the chapter will present some of the findings from the multiple sources of evidence collected. We will organize the results under the inferences set out by Knoch and Elder (2013) and have summarized the warrants and evidence in a table for each inference at the beginning of each section. Below each table, we describe the different sources of evidence in more detail and present the results for each.

4.1 Evaluation

Table 2.1 summarizes the three key warrants underlying the Evaluation inference. Evidence collected to find backing for each warrant is summarized in the final column.

To find backing for the first warrant in Table 2.1, the statistical properties of the PAAL were evaluated in the original trial as well as during the development of subsequent versions. Table 2.2 summarizes the Cronbach alpha results, which are all fairly consistent across the four versions. We also found a consistent spread of candidate abilities between new versions and the prototype version and a good spread of item difficulty.

Table 2.1 Warrants and related evidence for the Evaluation inference

Evaluation	
Warrants	Evidence
1. The psychometric properties of the test are adequate	Psychometric properties of the test as reported in the initial development report (Elder and Knoch 2009) and subsequent development reports (Knoch 2010a, b, 2011)
2. Test administration conditions are clearly articulated and appropriate	Responses to feedback questionnaires from the small trial and full trial (Knoch and O'Hagan 2014)
3. Instructions and tasks are clear to all test takers	Responses to feedback questionnaires and focus groups from the full trial (Knoch and O'Hagan 2014)

Table 2.2 Reliability estimates for the PAAL test versions

	Text completion (k = 75)	Speed reading (k = 75)	Combined screening (k = 150)	Writing
Version 1 (prototype)	.95	.96	.97	.883
Version 2	.92	.96	.97	n/a
Version 3	.93	.98	.98	n/a
Version 4	.95	.97	.98	n/a

The reliability statistics for the writing task used in this trial were also within acceptable limits for rater scored writing tasks, as was the case for previous versions (Elder et al. 2009).

To examine the second warrant in Table 2.1, we scrutinized the responses from the feedback questionnaires from the small and the full trial. The small trial of the online capabilities showed that there were several technical issues that needed to be dealt with before the test could be used for the full trial. For example, slow loading time tended to be an issue and some participants had experiences of the site 'crashing' or losing their connection with the site. The trial also indicated the need to further explore functionality to enable auto-correction features of some browsers to be disabled and adjustments to be made to font size for small screen users. In addition, feedback from trial participants indicated that fine-tuning of the test-taker interface was required. For example, some participants reported problems with the visibility/position of the on-screen timer and with the size of the text box for the writing task. The results of the trial further showed that there were variations in functionality across different platforms and devices (most notably, the iPad). Following this trial of the online capabilities of the system, a number of technical changes were made to the online system before the full trial was conducted.

Overall, the findings of the student survey and focus groups conducted following the full trial indicated that students' experiences of the online testing system were mostly positive in terms of accessibility of the website, clarity of the task instructions, and timely receipt of their report. Few students reported any technical problems, although there were a small number of students who found the system 'laggy', or slow to respond to keystrokes, and some reported that they had lost their internet

connection during the assessment. Overall, the purpose of the assessment and the benefits of taking it were clear to participating students and almost all of them appreciated being able to take the assessment from home in their own time.

The final warrant investigates whether students understood all the instructions and whether the task demands were clear. The questionnaire results from the two trials show that the students commented positively about these two areas.

In sum, the Evaluation inference was generally supported by the data collected from the different sources. The statistical properties of the PAAL were excellent, and the administration conditions suited the students and were adequate for the purpose, despite a few smaller technical problems which may have been caused by the internet rather than the PAAL software. The task demands and task instructions seemed clear to the test takers.

4.2 Generalizability

Table 2.3 lists the key warrants and associated supporting evidence we will draw on in our discussion of the Generalizability inference.

The first warrant supporting the Generalizability inference is that different test forms are parallel in design. The PAAL currently has four parallel forms or versions (and two more are nearly completed), all of which have been based on the same specification document. The psychometric results from the development of Versions 2, 3 and 4 show that each of these closely resembles the prototype version (Version 1).

The second warrant is that appropriate equating methods are used to ensure equivalence of test forms. The development reports of Versions 2, 3 and 4 outline the statistical equating methods that have been used to ensure equivalence in the meaning of test scores. Each time, a new version was trialed together with the anchor version and Rasch analysis was used to statistically equate the two versions.

Table 2.3 Warrants and related evidence for the Generalizability inference

Generalizability	
Warrants	Evidence
1. Different test forms are parallel in design	Review of test features and statistical evidence from reports of the development of parallel versions (Knoch 2010a, b, 2011)
2. Appropriate equating procedures are used to ensure equivalent difficulty across test forms	Review of equating evidence from reports of the development of parallel versions (Knoch 2010a, b, 2011)
3. Sufficient tasks are included to provide stable estimates of test taker ability	Psychometric properties of the test as reported in the initial development report (Elder and Knoch 2009) and subsequent development reports (Knoch 2010a, b, 2011)
4. Test administration conditions are consistent	Discussion of test delivery and results from the survey of the full trial (Knoch and O'Hagan 2014)

Statistically equating the writing tasks is more difficult as no suitable anchor items are available and only one writing task is included. However, the developers of the writing task attempt to closely stay true to the test specifications and small trials of new writing versions are carefully evaluated by a team of test developers to ensure they are as equivalent in design as possible and are eliciting assessable samples of writing performance from test candidates. Successive administrations of writing tasks for the DELA (from which the PAAL writing task is drawn) have shown stable estimates over different test versions as noted above.

The third warrant is that sufficient tasks are included to arrive at stable indicators of candidate performance. Each PAAL has 150 items, 25 for each of the three texts which make up the C-test and 75 in the cloze elide, as well as one writing task. As the PAAL is a screening test, the duration of 1 h is already at the upper limit of an acceptable amount of administration time. It is therefore practically impossible to add any more tasks. However, the trials have shown that the test results are fairly reliable indicators of test performance, with students being classified into the same categories when taking two parallel forms of the test.

The final warrant supporting generalizability relates to the consistency of the test administration. As students can take the test in their own time at a place of their choosing, it is likely that the conditions are not absolutely consistent. For example, a student might choose to take the test in a student computer laboratory that is not entirely free of noise, or at home in quiet conditions. However, due to the low stakes of the test, any differences in test taking conditions are probably not of great concern. Due to the fact that the test is computer-delivered, the timing and visual presentation of the test items are likely to be the same for all students.

By and large, it seems that the Generalizability inference is supported by the evidence collected from our trials.

4.3 Explanation and Extrapolation

Table 2.4 presents the warrants underlying the Explanation and Extrapolation inferences as well as the evidence we have collected.

The first warrant states that test takers' performance on the PAAL relates to their performance on other assessments of academic language proficiency. During the development of the prototype version of the PAAL, the cohort of students from the University of Melbourne had already taken the Diagnostic English Language Assessment (DELA) and their results could therefore be compared directly with their performance on the PAAL.

Table 2.5 presents the correlational results of the two PELA tests. It can be seen that overall screening test results correlated significantly with both the DELA overall raw scores and the DELA scaled scores.

The second warrant states that the scoring rubric captures relevant aspects of performance. The scoring rubric used to rate the writing performances has been developed on the basis of test developers' intuitions from their experience in EAP

Table 2.4 Warrants and related evidence for the Explanation and Extrapolation inferences

Explanation and Extrapolation	
Warrants	Evidence
1. Performance on the PELA relates to performance on other assessments of academic language proficiency	Correlational results from the development report (Elder and Knoch 2009)
2. Scoring criteria and rubrics capture relevant aspects of performance	Review of the literature on academic writing
3. Test results are good predictors of language performance in the academic domain	No data collected
4. Characteristics of test tasks are similar to those required of students in the academic domain (and those in the language development courses students are placed in)	No data collected
5. Linguistic knowledge, processes, and strategies employed by test takers are in line with theoretically informed expectations and observations of what is required in the corresponding academic context	No data collected
6. Tasks do not unfairly favor certain groups of test takers	No data collected

Table 2.5 DELA/AEST correlations (N = 156)

	C-test	Cloze elide	Screening total
DELA average (raw scores)	.772**	.721**	.809**
DELA average (scaled scores)	.775**	.699**	.797**

*p < .05, **p < .01

contexts as well as on a careful review of current practice in assessing writing in the academic domain. The criteria on the scale (organization and style, content and form) are commonly used in the assessment of academic writing and the level descriptions have been refined over the years to assist raters in better differentiating between candidates. Due to the nature of the task and the time limit of 30 min, the writing task captures only a limited sample from the candidates and there is no criterion specifically measuring the use of input reading material (which is in any case limited on this task – consisting only of a series of dot-point statements giving ideas for and against the proposition around which the argument is to be formulated). The ability to integrate reading input in writing is of course an important skill in academic writing but the benefits of assessing this ability needed to be weighed against the costs of devising and rating a more elaborate and time-intensive task involving extensive reading input. Although the task and its scoring rubric may somewhat under-represent the academic writing construct, the writing rating scale goes at least some way towards measuring relevant writing skills for the academic domain.

The third warrant states that test scores are good predictors of performance in the academic domain. No data in support of this warrant was collected as part of this study; however, an unpublished internal report (Group 2012) examining the relationship of the Diagnostic English Language Assessment (DELA) and students'

performance in their first year (as measured through WAMs[3]) shows that the DELA (which correlates strongly with the PAAL) is a very strong predictor of WAMs. The study clearly shows that a higher score on DELA is associated with higher WAMs and that a higher DELA score is associated with a lower risk of failing.

The fourth warrant states that the task types in the PELA are similar to those required of students in the academic domain. In the case of the PAAL, the test designers set out to develop a screening test which would be automatically scored and practical for test takers. It was therefore not possible to closely model the kinds of tasks test takers undertake in the academic domain (e.g. listening to a lecture and taking notes). However, the tasks chosen were shown to be very good predictors of the scores test takers receive on the more direct language tasks included in the DELA and it was therefore assumed that these indirect tasks could be used as surrogates. Similarly, warrant five sets out that the test takers' cognitive processes would be similar when taking the PELA and when completing tasks in the academic domain. Again, due to the very nature of the test tasks chosen, backing for this warrant might be difficult to collect. However, studies investigating the cognitive processes of test takers completing indirect tasks such as C-tests and cloze elide (e.g. Matsumura 2009) have shown that test takers draw on a very wide range of linguistic knowledge to complete these tasks, including lexical, grammatical, lexico-grammatical, syntactic and textual knowledge.

As for the final warrant, potential evidence has yet to be gathered from a larger test population encompassing students from different backgrounds, including native-English speaking (NES) and non-native speaking (NNES) students, and those in university Foundation courses, who may be from low literacy backgrounds or have experienced interrupted schooling. A previous study by Elder, McNamara and Congdon (2003) in relation to the DELNA screening component at Auckland would suggest that such biases may affect performance on certain items but do not threaten the validity of the test overall. Nevertheless, the warrant of absence of bias needs to be tested for this new instrument.

In sum, it would seem that the warrants for which evidence is available are reasonably well supported, with the caveat that the scope and screening function of the PAAL inevitably limits its capacity to fully represent the academic language domain.

4.4 Decisions

Table 2.6 sets out the warrants and associated evidence for the Decisions inference.

[3] WAM (weighted average mark) scores are the average mean results for students' first year course grades. The results of this in-house study are from an unpublished report undertaken for the University of Melbourne English Language Development Advisory Group committee which oversees the English language policy of the University of Melbourne.

Table 2.6 Warrants and related evidence for the Decisions inference

Decisions	
Warrants	Evidence
1. Students are correctly categorized based on their test scores	Review of standard-setting activities
2. The test results include feedback on test performance and a recommendation	Evidence from the questionnaires and focus groups of the full trial
3. The recommendation is closely linked to on-campus support	Review of institutional policy and evidence from focus groups of the full trial (Knoch and O'Hagan 2014)
4. Assessment results are distributed in a timely manner	Review of test documentation and evidence from focus groups of the full trial
5. The test results are available to all relevant stakeholders	Review of test procedures
6. Test users understand the meaning and intended use of the scores	Evidence from questionnaires and focus groups of the full trial

The first warrant in the Decision inference states that the students are categorized correctly based on their test score. Finding backing for this warrant involved two standard-setting activities. The first was conducted as part of the development of the prototype of the PAAL. A ROC (Receiver Operating Characteristics curve) analysis, a technique for setting standards, was used to establish optimum cut-scores or thresholds on the screening components of the test (c-test and cloze elide). A number of alternative cut-scores were proposed using either a specified DELA Writing score or an overall DELA score (representing the average of reading and listening and writing performance) as the criterion for identification of students as linguistically at risk. While these different cut-scores vary in sensitivity and specificity (see Elder and Knoch 2009 for an explanation of these terms), they are all acceptably accurate as predictors, given the relatively low stakes nature of the PAAL. Moreover, the level of classification accuracy can be improved through the use of the writing score to assist in decisions about borderline cases.

A further standard-setting exercise was conducted in preparation for the full trial. To set the cut-scores for the three result categories outlined in the previous section (i.e. 'proficient', 'borderline', 'at risk'), a standard- setting exercise was conducted with a team of trained raters at the Language Testing Research Centre using the writing scripts from the small trial. All 50 writing scripts were evaluated by the raters individually, evaluations were compared, and rating decisions moderated through discussion until raters were calibrated with each other and agreement was reached on the placement of each script in one of three proficiency groups: high, medium or low. To arrive at the two cut-scores needed, i.e. between 'proficient' and 'borderline', and between 'borderline' and 'at risk', we used the analytic judgement method (Plake and Hambleton 2001), a statistical technique for identifying the best possible point for the cut-score. Based on this, the cut-scores from the development trial were slightly shifted.

The second warrant states that test takers receive feedback on their performance and a recommendation. The feedback component following the PAAL is minimal, a fact that was criticized by the participants in the full trial. Students expressed disappointment with the results statement given in the report, describing it as somewhat generic and lacking in detail. Students in general would have preferred more diagnostic feedback to guide their future learning. Many also indicated they would have liked to discuss their report with an advisor to better understand their results, and to learn more about support opportunities.

Concerns were also raised about the vagueness of the support recommendation given in the report, with many students wanting a clearer directive for what was required of them. Many students stated they would have liked to receive follow up advice on how to act on the recommendation, with many reporting that they did not know how to access an advisor, or that they had received advice but it had not met their expectations.

The third warrant states that the recommendation is closely linked to on-campus support. Availability of appropriate support was identified as a problem by students who felt that offerings were not suited to their proficiency, level of study or academic discipline, or were otherwise not appropriate to their needs. Some students were also concerned that, in accessing the recommended support, they would incur costs additional to their course fees. Where a credit-bearing course was recommended, students expressed concerns about the implications for their academic record of failing the course.

The fourth warrant states that the assessment results are distributed in a timely manner. This was the case during the full trial, with results being distributed within 1–2 days of a student taking the assessment. Accordingly, in the evaluation of the full trial, students commented positively on the timely manner in which the results were distributed.

The next warrant relates to the availability of assessment results. During the full trial implementation, all accessible stakeholders were made aware of the fact that assessment results could be requested from the Language Testing Research Centre. Students were sent their results as soon as possible, and the Student Centres of the two cohorts were regularly updated with spreadsheets of the results. Unfortunately, because of the size of the cohorts, it was not possible to identify lecturers who would be responsible for teaching the students in question, and therefore lecturers may not have been aware of the fact that they could request the results.

The final warrant states that test users understand the meaning and the intended uses of the scores. The results of the evaluation of the full trial indicate that this was not an issue, at least from the test-taker perspective. The purpose of the assessment and the benefits of taking it were generally clear to students and students in the focus groups indicated that they all understood that the test was intended to be helpful, that it was important but not compulsory and that it did not affect grades.

Overall, the Decisions inference was only partially supported. We could find support for some aspects, including the categorization of the test takers, the expedient handling of test scores and that test takers generally understood their meaning. No data was collected from other test users, however, so the extent to which they

understood the meanings and intended uses of the scores requires further investigation. Other aspects relating to the feedback profile, the recommendation and the close link to on-campus support were not supported.

4.5 Consequences

Table 2.7 outlines the warrants of and associated evidence for the Consequences inference.

The first warrant underlying the Consequences inference states that all targeted test takers take the test, since this was the idea behind the streamlined test design (which was designed to be administered universally to all new students). During the full trial it became evident that institutional policy at the University of Melbourne makes it impossible to mandate such an assessment because it goes beyond content course requirements. Of the undergraduate Bachelor of Commerce cohort, only 110 students out of 310 (35%) took the assessment. The numbers were even lower for the Master of Engineering cohort, where only 60 students out of 491 (12%) of students took the PAAL. Evidence from the focus groups also showed that students differed in their understandings of whom the PAAL is for, with many believing it to be intended for 'international' students only. There was overall no sense among students that the assessment was meant to be universally administered.

The next warrant states that test takers' perceptions of the assessment are positive and that they find the assessment useful. The data from the full trial, some findings of which have already been reported under the Decision inference above, show mixed results. Students were generally positive about the ease of the test administration and the information provided prior to taking the test, but were less positive about the level of feedback provided and the follow-up support options available. Therefore, this warrant is only partially supported.

The next warrant states that the feedback from the assessment is useful and directly informs future learning. It is clear from the data from the full trial that the

Table 2.7 Warrants and related evidence for the Consequences inference

Consequences	
Warrants	Evidence
1. All targeted test takers sit for the test	Evidence from the full trial
2. The test does not result in any stigma or disadvantage for students	Evidence from the questionnaires and focus groups
3. Test takers' perceptions of the test and its usefulness are positive	Evidence from the questionnaires and focus groups
4. The feedback from the test is useful and directly informs future learning	Evidence from the questionnaires and focus groups
5. Students act on the test recommendation	Evidence from the questionnaires, focus groups of the full trial and follow-up correspondence with Student Centres

students did not find the feedback from the assessment particularly useful; however, it is important to remember that the PAAL is intended as a screening test, which is designed to identify students deemed to be at risk in minimum time and with minimum financial expenditure. When the test was designed, it was clear that no detailed feedback would be possible due to financial as well as practical limitations. While the analytically scored writing task potentially allowed for more detailed feedback, the resources to prepare this feedback were not available for a large cohort of students such as the one that participated in the full trial. More suitable online or on-campus support options would also have improved the chances of this warrant being supported. Unfortunately, offering more varied support provisions is costly and, in the current climate of cost-savings, probably not a viable option in the near future.

The final warrant states that students act on the score recommendation. Approximately 15 % of students taking the PAAL as part of the full trial were grouped into the 'at risk' group. It is not clear how many of these students acted upon the recommendation provided to them by enrolling in a relevant English language support course but historically the compliance rates at the University of Melbourne have been low. We suspect that the same would apply to this cohort for many reasons, including the limited array of support options and the lack of any institutional incentive or requirement to take such courses.

Overall, it can be argued that the Consequences inference was either not supported by the data collected in the full trial or that the relevant evidence was lacking.

5 Discussion and Conclusion

The chapter has described a new type of PELA instrument, which builds on previous models of PELA adopted in the Australian and New Zealand context. The online PAAL, taking just 1 h to administer, was designed to be quick and efficient enough to be taken by a large and disparate population of students immediately following their admission to the university, for the purpose of flagging those who might face difficulties with academic English and identifying the nature of their English development needs. Various types of validity evidence associated with the PAAL have been presented, using an argument-based framework for PELA validation previously explicated by the first two authors (Knoch and Elder 2013) and drawing on data from a series of trials.

The argument-based framework identifies the inferences that underlie validity claims about a test, and the warrants associated with each inference. The first is the *Evaluation* inference with its warrants of statistical robustness, appropriate test administration conditions and clarity (for test-takers) of tasks and instructions. These warrants were generally supported by the different sources of evidence collected, with an item analysis of each test component yielding excellent reliability statistics, the writing rater reliability being within expected limits, and feedback from test takers revealing that instructions were clear and tasks generally

manageable, apart from a few remediable technical issues associated with the online testing platform.

Warrants that were tested in relation to the second, *Generalizability*, inference were that different forms of the test were parallel in design and statistically comparable in level of difficulty, that there were sufficient tasks or items to provide stable estimates of ability and that test administration conditions were consistent. Results reported above indicate that different forms of the test were comparable, both in content and difficulty and that candidates were sorted into the same categories, regardless of the version they took. The online delivery of the PAAL moreover ensures consistency in the way test tasks are presented to candidates and in the time allowed for task performance.

As for the third *Explanation and Extrapolation* inference, which has to do with the test's claims to be tapping relevant language abilities, correlational evidence from the development trial showed a strong relationship between the PAAL scores for Parts A and B and the more time-intensive listening and reading items of the previously validated DELA, which had been administered concurrently to trial candidates. The warrant that the writing criteria capture relevant aspects of the academic writing construct is supported by research undertaken at the design stage. The predictive power of the writing component test is also supported by in-house data collected at the University of Melbourne showing the strong predictive power of DELA scores in relation to WAM scores. Other warrants associated with this inference have yet to be tested, however, and it is acknowledged that the length of the test and the indirect nature of the screening tasks in Parts A and B somewhat constrain its capacity to capture the academic language ability construct.

The *Decision* inference, the fourth in the argument-based PELA framework, encompasses warrants relating to the categorization of students based on test scores and the way test results are reported and received. Here the evidence presented gives a mixed picture. Standard-setting procedures ensured that the test's capacity to classify candidates into different levels was defensible. The meaning and purpose of the testing procedure was well understood by test users and score reports were made available to them in a timely manner. However, test-taker feedback revealed some dissatisfaction with the level of detail provided in the score reports and with the advice given about further support – perhaps because the avenues for such support were indeed quite limited. The fact that feedback was gathered from only a portion of the potential test taker population may also be a factor in these reactions as it tends to be the more motivated students who participate in trials. Such students are more likely to engage with the testing experience and expect rewards from it, including a full description of their performance and associated advice.

Evidence supporting the warrants relating to the fifth, *Consequences*, inference is even more patchy. Although the PAAL is designed to be administered to all incoming students, participation in the testing was by no means universal in the Faculties selected for the trial. In addition, there were mixed feelings about the usefulness of the initiative in informing future learning, due partly to the limited diagnostic information provided in the score reports but, more importantly, to the lack of available support options linked to these reports. Whether many students in the 'at

risk' category actually acted on their recommendation to enroll in support courses is unclear, but past history at the University suggests this is unlikely.

In general then, it can be seen that while the design of the PAAL and the information it provides about students' needs appears sound and indeed an improvement on previous PELA models in terms of its efficiency, there are issues associated with its utilization that require attention. Most of these issues are related to the policy environment in which the various trials were implemented rather than to the nature of the test itself.

For such a test to achieve its purpose of enhancing students' chances of academic success by identifying their particular English learning needs (or the lack of any such need), it has to be embedded in a more enlightened university policy which places a premium on the provision of opportunities for English language development, makes these opportunities accessible to students from any discipline by offering appropriate advice about avenues for action, and makes the consequences of inaction plain to test users, whether by mandating the test and enforcing its recommendations or through individual post-test counseling and follow-up tracking and monitoring of students. While the resources required to implement a fully-fledged language development policy alongside the test are considerable, the expenditure may well pay off in terms of student retention and outcomes which in turn would contribute to the reputation of the institution concerned. As well as suggesting such directions for policy reform, the findings of this study point to the necessity of construing the policy context as an integral dimension of validity, rather than merely as a set of external constraints.

References

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Bonanno, H., & Jones, J. (2007). *The MASUS procedure: Measuring the academic skills of university students – A diagnostic assessment*. Sydney: Learning Centre, University of Sydney. Available at: http://sydney.edu.au/stuserv/documents/learning_centre/MASUS.pdf

Dunworth, K. (2009). An investigation into post-entry English language assessment in Australian universities. *Journal of Academic Language & Learning, 3*(1), A1–A13.

Elder, C., & Knoch, U. (2009). *Report on the development and trial of the Academic English Screening Test (AEST)*. Melbourne: University of Melbourne.

Elder, C., & von Randow, J. (2008). Exploring the utility of a web-based English language screening tool. *Language Assessment Quarterly, 5*(3), 173–194.

Elder, C., McNamara, T., & Congdon, P. (2003). Rasch techniques for detecting bias in performance assessments: An example comparing the performance of native and non-native speakers on a test of academic English. *Journal of Applied Measurement, 4*(2), 181–197.

Elder, C., Knoch, U., & Zhang, R. (2009). Diagnosing the support needs of second language writers: Does the time allowance matter? *TESOL Quarterly, 43*(2), 351–359.

Ginther, A., & Elder, C. (2014). *A comparative investigation into understandings and uses of the TOEFL iBT test, the International English Language Testing System (Academic) Test, and the Pearson Test of English for Graduate Admissions in the United States and Australia: A case study of two university contexts* (ETS research report series, Vol. 2). Princeton: Educational Testing Service.

Group, E. L. D. A. (2012). *Performance in New Generation degrees and the University's policy on English language diagnostic assessment and support*. Melbourne: University of Melbourne.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17–64). Westport: American Council on Education/Praeger.

Klein-Braley, C. (1985). A cloze-up on the C-Test: A study in the construct validation of authentic tests. *Language Testing, 2*(1), 76–104.

Knoch, U. (2010a). *Development report of version 2 of the Academic English Screening Test (AEST)*. Melbourne: University of Melbourne.

Knoch, U. (2010b). *Development report of version 3 of the Academic English Screening Test*. Melbourne: University of Melbourne.

Knoch, U. (2011). *Development report of version 4 of the Academic English Screening Test*. Melbourne: University of Melbourne.

Knoch, U., & Elder, C. (2013). A framework for validating post-entry language assessments (PELAs). *Papers in Language Testing and Assessment, 2*(2), 1–19.

Knoch, U., & Elder, C. (2016). Post-entry English language assessments at university: How diagnostic are they? In V. Aryadoust & J. Fox (Eds.), *Trends in language assessment research and practice: The view from the Middle East and the Pacific Rim*. Newcastle-upon-Tyne: Cambridge Scholars Publishing.

Knoch, U., & O'Hagan, S. (2014). *Report on the trial implementation of the post-entry assessment of academic language (PAAL)*. Unpublished paper. Melbourne: University of Melbourne.

Knoch, U., Elder, C., & McNamara, T. (2012a). *Report on Feasibility Study of introducing the Academic English Screening Test (AEST) at the University of Melbourne*. Melbourne: University of Melbourne.

Knoch, U., O'Hagan, S., & Kim, H. (2012b). *Preparing the Academic English Screening Test (AEST) for computer delivery*. Melbourne: University of Melbourne.

Matsumura, N. (2009). *Towards identifying the construct of the cloze-elide test: A mixed-methods study.* Unpublished Masters thesis, University of Melbourne.

Plake, B., & Hambleton, R. (2001). The analytic judgement method for setting standards on complex performance assessments. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 283–312). Mahwah: Lawrence Erlbaum.

Read, J. (2015). *Assessing English proficiency for university study*. London: Palgrave Macmillan.

Chapter 3
Mitigating Risk: The Impact of a Diagnostic Assessment Procedure on the First-Year Experience in Engineering

Janna Fox, John Haggerty, and Natasha Artemeva

Abstract The global movement of students, the linguistic and cultural diversity of university classrooms, and mounting concerns about retention and program completion have prompted the increased use of post-entry diagnostic assessment, which identifies students at risk and provides them with early academic support. In this chapter we report on a multistage-evaluation mixed methods study, now in its sixth year, which is evaluating the impact of a diagnostic assessment procedure on the first-year experience, student engagement, achievement, and retention in an undergraduate engineering program. The diagnostic assessment procedure and concomitant student support are analyzed through the lens of Activity Theory, which views socio-cultural object-oriented human activity as mediated through the use of *tools*, both symbolic (e.g., language) and material (e.g., computers, pens). Changes in Activity Systems and their interrelationships are of central interest. In this chapter we report on changes resulting from modifications to the diagnostic assessment procedure that have increased its impact on the first-year experience by: (1) applying a disciplinary (rather than generic) assessment approach which was *fine grained* enough to trigger actionable academic support; (2) embedding the diagnostic assessment procedure within a required first-year engineering course, which increased the numbers of students who voluntarily sought support; and (3) paying increased attention to the development of social connections, which play an important role in student retention and success.

Keywords Activity theory • At-risk students • Diagnostic assessment • Engineering education • First-year experience • Higher education • Retention • Social connection

J. Fox (✉) • N. Artemeva
School of Linguistics and Language Studies, Carleton University, Ottawa, Canada
e-mail: Janna.Fox@carleton.ca; Natasha.Artemeva@carleton.ca

J. Haggerty
Department of Language and Literacy Education, University of British Columbia, Vancouver, Canada
e-mail: john.haggerty@alumni.ubc.ca

J. Read (ed.), *Post-admission Language Assessment of University Students*,
English Language Education 6, DOI 10.1007/978-3-319-39192-2_3

1 Introduction

Pre-admission language proficiency testing has become a ubiquitous requirement for second language applicants to English-medium universities. Over the years, test users have tended to mistakenly interpret high scores on language proficiency tests as evidence of academic readiness (see, Fox et al. 2016), but if language proficiency alone were sufficient for successful academic engagement in an English-medium university, all English-speaking students would be successful. Clearly this is not the case.

In Canada, alongside the trend to ever larger numbers of international students, decades of immigration have contributed to increasing cultural and linguistic diversity in university classrooms (Anderson 2015). It is estimated that up to half of the students in Canada's schools speak a language other than English or French[1] as their first language (Fox 2015). In response to increasing cultural and linguistic diversity and greater concern about retention, universities have been developing a wide array of support services to address student needs. Typically, these services are generic and centralized, often with special attention directed at first-year undergraduates. For example, at the university where this study took place there are a number of such services available to students, including a writing tutorial service and a math tutorial service, amongst others.

Statistics on success in university suggest that first-year undergraduate students are the most likely to drop out or *stop out* (i.e., leave university for a period of time but return later) during the first term/s of their university programs (Browne and Doyle 2010). For this reason, *the first-year experience* has become a focus of much of the literature on university retention (e.g., Browne and Doyle 2010; Tinto 1993). Although there is evidence that generic support programs are helpful to students (Cheng and Fox 2008; Fox et al. 2014), at the university which is the site of the present study, statistics on retention rates suggest that such services have had relatively little impact on retention in the engineering program. For example, from 2007 through 2012, this program lost on average 16 % of its students by the end of the first year; another 8 % after second year (a 24 % cumulative loss); and an additional 7 % after third year (a 31 % cumulative loss) (Office of Institutional Research and Planning 2014). Further, students within the engineering program were often taking the same course several times, and as a result, a number of them were not completing their programs in a timely manner. Many of the courses with the highest failure rates and repeat registrations were first- or second-year courses, which were part of the foundation or core engineering curriculum that all engineering students are required to complete successfully before specializing in an engineering discipline (e.g., civil, mechanical, electrical).

Within this context, and in light of mounting concerns in the Faculty of Engineering about retention, patterns of course enrollments (i.e., failure and

[1] English and French are official languages of Canada and serve as mediums of instruction in Canadian universities.

repeated registration(s) in the same course), and delays in program completion, a post-entry diagnostic assessment procedure was implemented in 2010 for entering undergraduate students. The intent of the assessment was to identify at-risk students early in their program and provide them with individualized pedagogical support to mitigate risk of failure. Engineering professors, Teaching Assistants (TAs) and other stakeholders reported that, in addition to issues with language and writing, a number of first-year students lacked threshold concepts in mathematics, misunderstood instructions for assignments, underestimated how much work they needed to do on an on-going basis so as not to fall behind, and so on. It was evident that more than English language proficiency was creating risk for entering undergraduates in engineering.

In this chapter we report on a mixed methods research study examining the impact of the diagnostic assessment procedure, now in its sixth year, and guided by the following research question: Does a diagnostic assessment procedure, combining assessment with individual pedagogical support, improve the experience, retention, and achievement of entering undergraduate engineering students? This chapter reports on the impact of the diagnostic procedure on the first-year experience.

Before discussing the study itself, in the section which follows we discuss the diagnostic assessment procedure that has been developed for this university context within the theoretical framework and empirical research that informed it. We also provide background on the challenges and concerns that have characterized its implementation.

2 Evolution of the Diagnostic Assessment Procedure

Huhta (2008) distinguishes formative assessment from diagnostic assessment, submitting that they are on two ends of an assessment continuum. He suggests that formative assessment is rooted in on-going curricular and classroom concerns for feedback, teaching, and learning in *practice*, whereas diagnostic assessment is informed by theory and theoretical models of language and learning. However, in our view it is their connection, rather than what distinguishes them, that is of central concern. We use the phrase *diagnostic assessment procedure* to underscore that diagnostic assessment and concomitant pedagogical intervention are inseparable. We argue that a diagnostic assessment procedure cannot be truly diagnostic unless it is linked to feedback, intervention, and support. Alderson (2007) suggests this in his own considerations of diagnostic assessment:

> … central to diagnosis must be the provision of usable feedback either to the learners themselves or to the diagnoser—the teacher, the curriculum designer, the textbook writer, and others. ... [T]he nature of [the] feedback, the extent to which it can directly or indirectly lead to improvements in performance or in eradicating the weaknesses identified, must be central to diagnostic test design. (p. 30)

In this study, the diagnostic assessment procedure has as its goal to: (1) identify entering students at-risk in the first year of their undergraduate engineering program; and (2) generate a useful learning profile (i.e., what Cai 2015 notes must lead to *actionable feedback*), which is linked to individually tailored (Fox 2009) and readily available academic support for the learning of a first-year engineering student.

The study is informed by sociocultural theory (Vygotsky 1987), which views knowledge as contextualized and learning as social (e.g., Artemeva and Fox 2014; Brown et al. 1989; Lave and Wenger 1991). From this perspective, both the assessment and the pedagogical interactions that it triggers are *situated* within and informed by the context and the community in which they occur (Lave and Wenger 1991). However, context is a complex and multi-layered construct, extending from micro to increasingly (infinitely) macro levels. Thus, early in the study we encountered what is known in the literature as the "frame problem" (Gee 2011a, p. 67, 2011b, p. 31), namely, to determine the degree of situatedness that would be most useful for the purposes of the diagnostic assessment procedure.

In the initial implementation of the diagnostic assessment, we operationalized what Read (2012) refers to as a general academic language proficiency construct, of relevance to university-level academic work. For our pilot in 2010, we drew three tasks from the Diagnostic English Language Needs Assessment (DELNA) test battery (see, Read 2008, 2012; or, the DELNA website: http://www.delna.auckland.ac.nz/). DELNA is a Post-Entry Language Assessment (PELA) procedure. As discussed in Read (2012), such PELA procedures operationalize the construct of academic English proficiency and draw on generic assessment materials of relevance across university faculties and programmes. Alderson (2007) argues that "diagnosis need not concern itself with authenticity and target situations, but rather needs to concentrate on identifying and isolating components of performance" (p. 29). Isolating key components of performance as the result of diagnosis provides essential information for structuring follow-up pedagogical support.

Of the three DELNA tasks that were used for the initial diagnostic assessment procedure, two were administered and marked by computer and tested academic vocabulary knowledge and reading (using multiple-choice and cloze-elide test formats). The third task tested academic writing with a graph interpretation task that asked test takers to write about data presented in a histogram. The task was marked by raters trained to use the DELNA rubric for academic writing. We drew two groups of raters, with language/writing studies or engineering backgrounds, and all were trained and certified through the DELNA on-line rater training system (Elder and von Randow 2008). However, while the language/writing raters tended to focus on language-related descriptors, the engineering raters tended to focus on content. Although inter-rater reliability was high (.94), there was a disjuncture between what the raters attended to in the DELNA rating scale and what they valued in the marking (Fox and Artemeva 2011). Further, some of the descriptors in the DELNA generic grid were not applicable to the engineering context (e.g., valuing length as opposed to concise expression; focusing on details to support an argument, rather than interpreting trends). The descriptors needed to map onto specific, actionable

pedagogical support summarized in the learning profile. Having descriptors which were not of relevance to engineering was unhelpful.

Following the research undertaken by Fox and Artemeva (2011), we recognized that considerations of disciplinarity were critical if interventions were to provide what Prior (1994) refers to as an "opportunity space for socialization into [the] discursive practices" (p. 489) of the discipline. As a result, we narrowed the *frame* to the context of first-year undergraduate engineering and began to operationalize an engineering-specific, *academic literacies construct* (Read 2015). The diagnostic assessment procedure aimed to create a safe support space which would allow new students (novices in the discipline) to begin to "display disciplinarity" (Prior 1994, p. 489). As Artemeva (2006) points out, novices typically go through a fairly slow process of acculturation before they can communicate what they know in ways acceptable to a new discipline. A Centre, staffed by peer mentors who were drawn from the pool of raters, was set up to provide support which would help facilitate a new student's acculturation. This support space was much *safer* than the classroom, which is public, contested, and subject to grades or marks; safer, because the Centre was staffed by knowledgeable student mentors who were nonetheless external to courses and posed no threat.

3 Theoretical Framework

Our analysis of the diagnostic assessment procedure and concomitant student support discussed in this chapter are enriched through the use of Activity Theory (AT) (Engeström 1987; Engeström et al. 1999; Leont'ev 1981). Vygotsky (1987) argued that object-oriented human activity is always mediated through the use of signs and symbols (e.g., language, texts). One of Vygotsky's students, Leont'ev (1981), later posited that *collective* human activity is mediated not only by symbolic, but also by material tools (e.g., pens, paper). Leont'ev (1981) represented human activity as a triadic structure, often depicted as a triangle, which includes a subject (i.e., human actors), an object (i.e., something or someone that is acted upon), and symbolic or material tools which mediate the activity. The subject has a motive for acting upon the object in order to reach an outcome. AT was further developed by Engeström (1987), who observed that "object-oriented, collective and culturally mediated human activity" (Engeström and Miettinen 1999, p. 9) is best modelled as an *activity system* (e.g., see Figs. 3.1 and 3.2). In other words, "an activity system comes into existence when there is a certain need . . . that can be satisfied by a certain activity" (Artemeva 2006, p. 44). As Engeström noted, multiple activity systems interact over time and space. In the present chapter, we consider the diagnostic assessment procedure as comprising two interrelated activity systems (Fig. 3.1): a diagnostic assessment activity and a pedagogical support activity.

The diagnostic assessment procedure prompts or instigates the development of the activity system of the undergraduate engineering student (Fig. 3.2), who voluntarily seeks support from peer mentors within the Centre in order to improve his or her academic achievement (i.e., grades, performance).

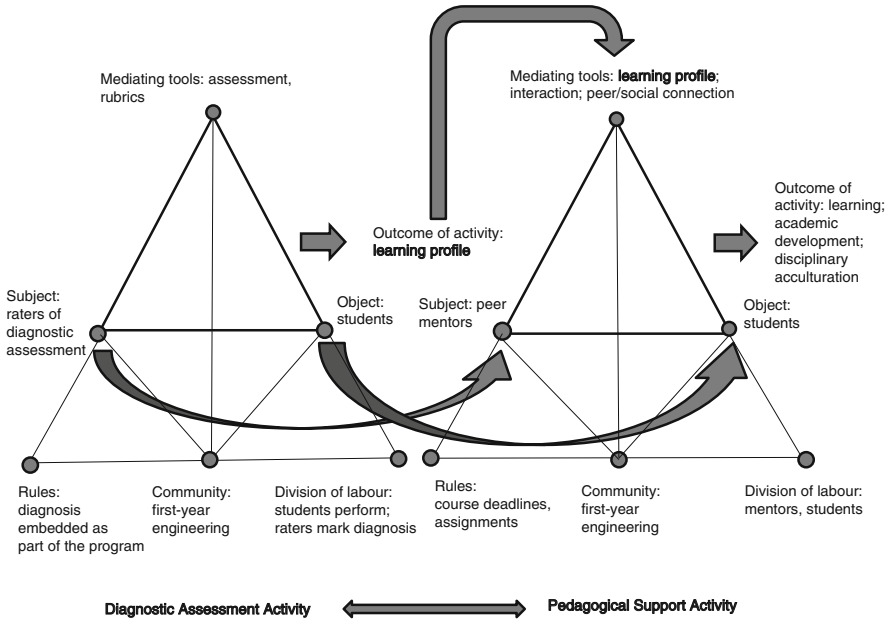


Fig. 3.1 Activity systems of the diagnostic assessment procedure in undergraduate engineering

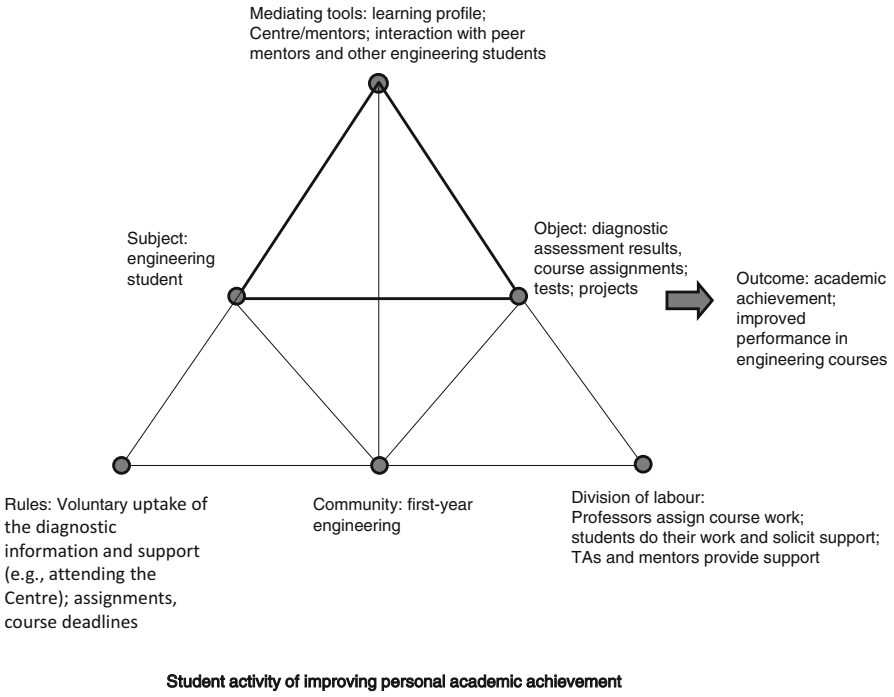


Fig. 3.2 The activity system of an undergraduate engineering student

It is important to note that the actors/subjects in the activity systems of the diagnostic assessment procedure (Fig. 3.1) are not the students themselves. As indicated above, the raters in the diagnostic assessment activity system become peer mentors in the pedagogical support activity. However, the students who use the information provided to them in the learning profile and seek additional feedback or pedagogical support bring into play another activity system (Fig. 3.2), triggered by the activities depicted in Fig. 3.1.

In our study, the activities presented in Figs. 3.1 and 3.2 are all situated within the community of undergraduate engineering, and, as is the case in any activity system (Engeström 1987), are inevitably characterized by developing tensions and contradictions. For example, there may be a contradiction between the mentors' and the students' activity systems because of contradictions in the motives for these activities. Mentors working with students within the Centre are typically motivated to support students' *long-term learning* in undergraduate engineering, whereas most students tend to have *shorter-term* motives in mind, such as getting a good grade on an assignment, clarifying instructions, or unpacking what went wrong on a test. According to AT, such tensions or contradictions between the mentors' activity system and the students' activity system are the site for potential change and development (Engeström 1987). Over time, a student's needs evolve and the student's motive for activity may gradually approach the motive of the mentors' activity. One of the primary goals of this study is to find evidence that the activity systems in Figs. 3.1 and 3.2 are aligning to ever greater degrees as motives become increasingly interrelated. Evidence that the activity systems of mentors and students are aligning will be drawn from students' increasing capability, and awareness of what works, why, and how best to communicate knowledge and understanding to others within their community of undergraduate engineering as an outcome of their use of the Centre. Increased capability and awareness enhance the first-year experience (Artemeva and Fox 2010; Scanlon et al. 2007), and ultimately influence academic success.

The diagnostic assessment procedure is also informed by empirical research which is consistent with the AT perspective described above. This research has investigated undergraduates' learning in engineering (e.g., Artemeva 2008, 2011; Artemeva and Fox 2010), academic and social engagement (Fox et al. 2014), and the process of academic acculturation (Cheng and Fox 2008), as well as findings produced in successive stages of the study itself (e.g., Fox et al. 2016). We are analyzing longitudinal data on an on-going basis regarding the validity of the at-risk designation and the impact of pedagogical support (e.g., retention, academic success/failure, the voluntary use of pedagogical support), and using a case study approach to develop our understanding of the phenomenon of *risk* in first-year engineering.

Having provided a brief discussion of the theoretical framework and on-going empirical research that inform the diagnostic assessment procedure, in the section below we describe the broader university context within which the procedure is situated and the issues related to its development and implementation.

4 Implementation of the Diagnostic Assessment: A Complex Balancing Act

The development and implementation of the diagnostic assessment procedure described here may be best described as a complex and challenging balancing act (Fox and Haggerty 2014). In maintaining this balance, there have been issues and tensions (see Fig. 3.3) regarding:

1. Assessment quality (e.g., tasks, scoring rubrics, rater consistency);
2. Pedagogical support; and
3. Presentation and marketing of the diagnostic assessment to key stakeholders (e.g., students, administrators, faculty, TAs).

In the section below, we discuss each of these dimensions of concern in relation to some of the questions that we have attempted to address over the 6 years of development, administration, and implementation of the diagnostic assessment procedure. This is not a comprehensive list by any means; rather, it provides an overview of the key questions that we needed to answer with empirical evidence garnered from on-going research.

Responses to the following questions have guided decision-making with regard to the three dimensions of concern:

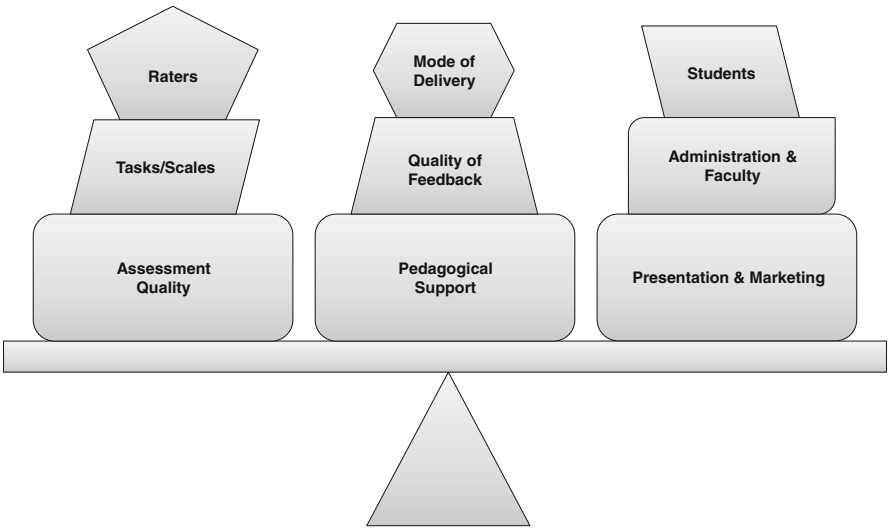


Fig. 3.3 Issues and tensions in diagnostic assessment: an ongoing balancing act

4.1 Assessment Quality

When a mandate for an assessment procedure had been established, and it had been determined that a test or testing procedures would best address the mandate, the critical first question is 'do we know what we are measuring?' (Alderson 2007, p. 21). As Alderson points out, "Above all, we need to clarify what we mean by diagnosis … and what we need to know in order to be able to develop useful diagnostic procedures" (p. 21). Assessment quality depends on having theoretically informed and evidence-driven responses to each of the following key questions:

4.1.1 Is the construct well-enough understood and defined to warrant its operational definition in the test?
4.1.2 Are the items and tasks meaningfully operationalizing the construct? Do they provide information that we can use to shape pedagogical support?
4.1.3 Is the rating scale sufficiently detailed to provide useful diagnostic information?
4.1.4 Are the raters consistent in their interpretation of the scale?
4.1.5 Does the rating scale and the resulting score or scores provide sufficient information to trigger specific pedagogical interventions?

4.2 Pedagogical Support

Once the results of a diagnostic assessment are available, there are considerable challenges in identifying the most effective type of pedagogical support to provide. The other contributors to this volume have identified the many different approaches taken to provide academic support for students, especially for those who are identified as being at-risk. For example, in some contexts academic counsellors are assigned to meet with individual students and provide advice specific to their needs (e.g. Read 2008). In other contexts, a diagnosis of *at-risk* triggers a required course or series of workshops.

Determining which pedagogical responses will meet the needs of the greatest number of students and have the most impact on their academic success necessitates a considerable amount of empirical research, trial, and (unfortunately) error. Tensions occur, however, because there are always tradeoffs between what is optimal support and what is practical and possible in a given context (see Sect. 4.3 below).

Research over time helps to identify the best means of support. Only with time is sufficient evidence accumulated to support pedagogical decisions and provide evidence-driven responses to the following questions.

4.2.1 Should follow-up pedagogical support for students be mandatory? If so, what should students be required to do?
4.2.2 Should such support be voluntary? If so, how can the greatest numbers of students be encouraged to seek support? How can students at-risk be reached?

4.2.3 What type of support should be provided, for which students, when, how, and by whom?
4.2.4 How precisely should information about test performance and pedagogical support be communicated to the test taker/student? What information should be included?
4.2.5 How do we assess the effectiveness of the pedagogical support? When should it begin? How often? For what duration?
4.2.6 What on-going research is required to evaluate the impact of pedagogical interventions? Is there evidence of a change in students' engagement and their first-year experience?
4.2.7 What evidence do we need to collect in order to demonstrate that interventions are working to increase overall academic retention and success?

4.3 Presentation and Marketing

Diagnostic assessment and subsequent pedagogical responses are time-consuming and costly. One of the key tensions in considerations of diagnostic assessment procedures is the need to provide evidence that additional cost is warranted by the positive benefits that result.

Presentation is critical to persuading university administrators who control budgets that offering a diagnostic assessment to students early in their undergraduate programs will promote retention and academic success. From the beginning, it is important to collect evidence on an on-going basis to demonstrate how diagnosis and intervention are making a difference.

In some contexts, students at-risk are required to undertake a course, participate in workshops, or receive counselling as a result of the diagnosis. Although some universities may provide the necessary funding to cover these additional costs, in many instances students who are required to take an extra course must pay for it themselves or pay a small fee (in the case of additional, mandatory workshops). In the context of the diagnostic assessment procedure which is the focus of the present chapter, students were not required to use the diagnostic information. Students' use or follow-up was entirely voluntary. Thus, presentation and marketing of the usefulness of the diagnostic assessment procedure to students was a key concern and led to many challenges which were addressed through evidence-driven responses to the following key questions:

4.3.1 How can we best encourage students to follow-up on the diagnostic information provided by the assessment?
4.3.2 What evidence should be collected in order to persuade administrators that the assessment procedure is having an impact on retention and academic success?
4.3.3 How, when, and where should pedagogical support be delivered?
4.3.4 Who should provide pedagogical support? What funding is available for this cost?

4.3.5 How should we recruit and train personnel to provide effective pedagogical support?
4.3.6 Who should monitor their effectiveness?
4.3.7 How clear is the tradeoff between cost and benefit?
4.3.8 How can we tailor the diagnostic assessment procedure and pedagogical follow-up to maximize practicality, cost-effectiveness, and impact?

Given the complex balancing act of developing and implementing a diagnostic assessment procedure, we have limited our discussion in the remainder of this chapter to two changes that have occurred since the assessment was first introduced in 2010. These evidence-driven changes were perhaps the most significant in improving the quality of the assessment itself and increasing the overall impact of the assessment procedure. The two changes that were implemented in 2014 are:

- Embedding diagnostic assessment within a mandatory, introductory engineering course; and,
- Providing on-going pedagogical support during the full academic year through the establishment of a permanent support Centre for engineering students.

These changes were introduced as a result of research studies, occurring at different phases and multiple stages of development and implementation of the procedure, informed by the theoretical framework, and guided by the questions listed above and the overall research question: Does a diagnostic assessment procedure, combining assessment with individual pedagogical support, improve the first-year experience, achievement, and retention of undergraduate engineering students? If so, in what ways, how, and for whom? In this chapter we focus on evidence which relates the two changes in 2014 to differences in the first-year experience.

5 The Current Study: Two Significant Changes

5.1 Method

The two key changes to the diagnostic assessment procedure, namely, embedding it in a required course and providing a permanent Centre for support, were supported by initial results of the longitudinal mixed-methods study which is exploring the effectiveness of this diagnostic assessment procedure by means of a *multistage-evaluation design* (Creswell 2015). As Creswell notes, "[t]he intent of the multistage evaluation design is to conduct a study over time that evaluates the success of a program or activities implemented into a setting" (p. 47). It is multistage in that each phase of research may, in effect, constitute many stages (or studies). These stages may be qualitative (QUAL or qual, depending upon dominance), quantitative (QUAN or quan, depending upon dominance) or mixed methods, but taken together they share a common purpose. Figure 3.4 provides an overview of the research design and includes information on the studies undertaken within phases of the

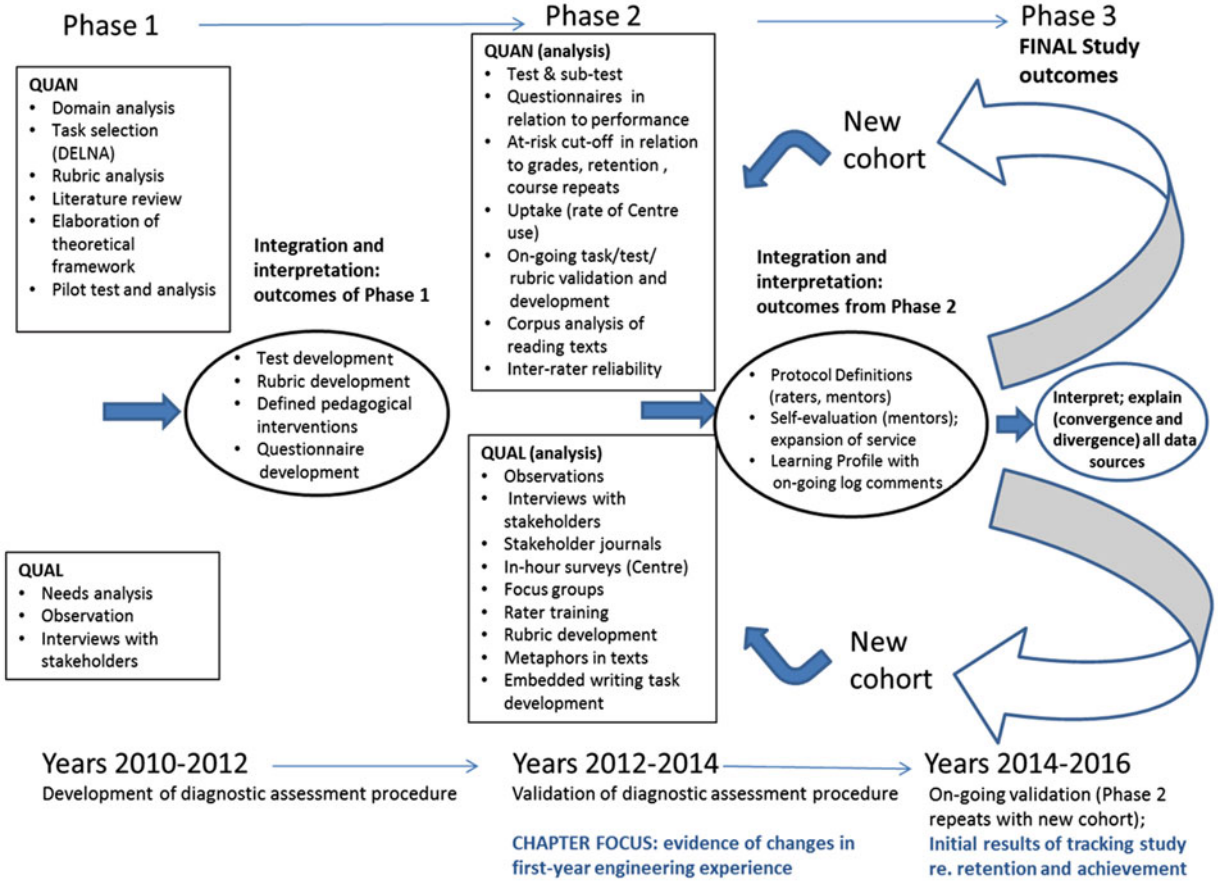


Fig. 3.4 Mixed methods design of the diagnostic assessment procedure

research; how the qualitative and quantitative strands were integrated; and the relationship of one phase of the research to the next from 2010 to 2016.

Engaging in what is essentially a 'program-of-research' approach is both complex and time-consuming. Findings are reported concurrently as part of the overall research design. For example, within the umbrella of the shared purpose for this longitudinal study, Fox et al. (2016) report on a Phase 1 research project investigating the development of a writing rubric, which combines generic language assessment with the specific requirements of writing for undergraduate engineering purposes. As discussed above, the new rubric was the result of tensions in the diagnostic assessment procedure activities (Fig. 3.1) in Phase 1 of the study. The generic DELNA grid and subsequent learning profile contradicted some of the expectations of engineering writing (e.g., awarded points for length, anticipated an introduction and conclusion, did not value point form lists). Increasing the engineering relevance of the rubric also increased the diagnostic feedback potential of the assessment procedure and triggered differential pedagogical interventions. In other words, tensions in the original activity system led to the development of a more advanced activity system (Engeström 1987), which better identifies and isolates components of performance (Alderson 2007). In another Phase 1 project, McLeod (2012) explored the usefulness of diagnostic feedback in a case study which tracked the academic acculturation (Artemeva 2008; Cheng and Fox 2008) of a first-year undergraduate engineering student over the first year of her engineering program. She documented the interaction of diagnostic assessment feedback, pedagogical support, and the student's exceptional social skill in managing risk, as the student navigated through the challenges of first-year engineering. McLeod's work further supports the contention that social connections within and across a new disciplinary community directly contribute to a student's academic success.

Another component of the multistage evaluation design includes a qualitative case study in Phase 2 with academic success in engineering being the phenomenon of interest. The case study is drawing on semi-structured interviews with first-year and upper-year engineering students, TAs, professors, peer mentors, administrators, etc. to investigate the impact of the diagnostic assessment from different stakeholder perspectives. There is also a large-scale quantitative study in Phase 2 which is tracking the academic performance of students who were initially identified as being at-risk and compares their performance with their more academically skilled counterparts using various indicators of student academic success (e.g., course withdrawals, course completions and/or failures, drop-out rates, and use of the Centre which provides pedagogical support).

As indicated above, the design is recursive in that the analysis in Phase 2 will be repeated for a new cohort of entering engineering students and the results will further develop the protocols designed to support their learning. The current study will be completed in 2016. On-going in-house research is now the mandate of the permanent Centre (see Fig. 3.4, Phase 3).

Table 3.1 Summary of participants by stakeholder group (2010–2015)

Stakeholder group	Pilot test	External to course	External to course	Embedded in course	Embedded in course
Year	2010–2011	2011–2012	2012–2013	2013–2014	2014–2015
Engineering students	160	489	518	971	1,014[a]
Professors/instructors	5	3	5	5	5
Peer mentors		11	5	7	11[b]
Trained raters	15	8	4	6	6
Administrators (project coordinator; dean; associate dean, etc.)	1	1	3	3	3
Total	181	512	535	992	1,039

[a]972 students were tested in September 2014; 42 were tested in January 2015
[b]3 were upper-year students in writing/language studies; 8 engineering

5.2 Participants

In large-scale mixed-methods studies which utilize a multistage-evaluation design, the numbers and types of participants tend to fluctuate continuously over time (Table 3.1).

As Creswell (2015) points out and as the name implies, this advanced type of mixed methods design is comprised of many stages within multiple phases, all of which collectively support a sustained line of inquiry, (e.g., needs assessment, development of a conceptual framework, field testing of prototypes). Each phase in the inquiry may feature "rigorous quantitative and qualitative research methods" (p. 3) or mixed methods, but the core characteristic of such a study is the *integration* (Fig. 3.4) of findings in relation to the overall intent of the research. What distinguishes a multistage evaluation design is that "integration consists of expanding one stage into other stages over time" (p. 7).

Although each stage of research responds to specific questions which dictate a particular sampling strategy (Creswell 2015), and the large-scale tracking study is considering the whole population of undergraduates in engineering, in Table 3.1 we provide an overview of participants from 2010 (when the initial assessment consisting of three DELNA tasks was first pilot tested) to 2015.

5.3 Findings and Discussion

As indicated in the introduction to this chapter, from the beginning there were two overarching concerns to address in evaluating the impact of the diagnostic assessment procedure within this engineering context:

- Development of a post-entry diagnostic assessment which would effectively identify undergraduate engineering students at-risk early in their first term of study; and

- Provision of effective and timely academic support for at-risk students, as well as any other first-year engineering students, who wanted to take advantage of the support being offered.

Initially, although students were directly encouraged to take the diagnostic assessment, their participation and use of the information provided by the assessment as well as follow-up feedback and post-entry support was *voluntary*. There were no punitive outcomes (e.g., placement in a remedial course; required attendance in workshops; reduction in course loads and/or demotion to part-time status). Students received feedback and advice on their diagnostic assessment results by email a week after completing the assessment. Their performance was confidential (neither their course professors nor the TAs assigned to their courses were informed of their results). The emails urged students to drop in for additional feedback at a special Centre to meet with other, upper year students (in engineering and writing/language studies) and get additional feedback, information, and advice on how to succeed in their engineering courses.

In 2014, two critical changes occurred in the delivery of the diagnostic assessment procedure which dramatically increased its impact. These two changes, which are the focus of the findings below, were the cumulative result of all previous stages of research, and as indicated above, have had to date the largest impact on the quality of the diagnostic assessment procedure.

Each of the key changes implemented in 2014 is discussed separately in relation to the findings which informed the changes.

5.3.1 Evidence in Support of Embedding the Diagnostic Assessment Procedure in a Mandatory First-Year Engineering Course

In Phase 1 of the study (2011–2012), 489 students (50 % of the first-year undergraduate engineering cohort) were assessed with three of DELNA's diagnostic tasks leased from the University of Auckland (DELNA's test developer). The DELNA tasks were administered during *orientation week* – the week which precedes the start of classes in a new academic year and introduces new students to the university. Students were informed of their results by email and invited on a voluntary basis to meet with peer mentors to receive pedagogical support during the first months of their engineering program. A Support Centre for engineering students was set up during the first 2 months of the 2011–2012 term. It was staffed by upper-year engineering and writing/language studies students who covered shifts in the Centre from Monday through Friday, and who had previously rated the DELNA writing sub-test.

During the 6 weeks the Centre was open, only 12 (2 %) of the students sought feedback on their diagnostic assessment results. The students tended to be those who were outstanding academically and would likely avail themselves of every opportunity to improve their academic success, or students who wanted information on an assignment. Only 3 of the 27 students who were identified as at-risk (11 %) visited the Centre for further feedback on their results and took advantage of the

pedagogical support made available. At the end of the academic year, ten of the at-risk group had dropped out or were failing; seven were borderline failures; and, ten were performing well (including the three at-risk students who had sought feedback).

In 2012–2013, 518 students (70 % of cohort) were assessed, but only 33 students (4 %) voluntarily followed-up on their results. However, there was evidence that three of these students remained in the engineering program because of early diagnosis and pedagogical intervention by mentors in the Centre. Learning of the success of these three students, the Dean of Engineering commented, "Retaining even two students pays for the expense of the entire academic assessment procedure."

In 2013–2014, the DELNA rating scale was adapted to better reflect the engineering context. This *hybrid* writing rubric (see Fox et al. 2016) improved the *grain size* or specificity of relevant information provided by the rubric and enhanced the quality of pedagogical interventions. Further, DELNA's graph interpretation task was adapted to represent the engineering context. Graphic interpretation is central to work as a student of engineering (and engineering as a profession as well). However, when the DELNA graph task was vetted with engineering faculty and students, they remarked that "engineers don't do histograms". This underscored the importance of disciplinarity (Prior 1994) in this diagnostic assessment procedure.

It became clear that many of the versions of the generic DELNA graph task were more suited for social science students than for engineering students, who most frequently interpret *trends* with line graphs. In order to refine the diagnostic feedback elicited by the assessment and shape pedagogical interventions to support students in engineering, it became essential that engineering content, tasks, and conventions be part of the assessment. Evidence suggested that the *frame* of general academic language proficiency (Read 2015) was too broad; decreasing the frame size and situating the diagnostic assessment procedure with engineering text, tasks, and expectations of performance also increased both the overall usefulness of the assessment (Bachman and Palmer 1996) as well as the relevance and specificity (grain size) of information included in the learning profiles of individual students. More specificity in learning profiles also increased the quality and appropriacy of interventions provided to students in support of their learning. From the perspective of Activity Theory, the mentors were increasingly able to address the students' motives: to use the learning profiles as a starting point; to mediate activity in relation to the students' motives to improve their performance in their engineering classes or achieve higher marks on a lab report or problem set. The increased relevance of the mentors' feedback and support suggests increased alignment between the activity systems of mentors and students (Figs. 3.1 and 3.2).

Domain analysis which investigated undergraduate engineering supported the view that engineering students might be at-risk due to more than academic language proficiency. For example, some students had gaps in their mathematics background while others had difficulty reading scientific texts. Still others faced challenges in written expression required for engineering (e.g., concise lab reports; collaborative or team writing projects). Importantly, a number of entering students, who were

deemed at-risk when the construct was refined to reflect requirements for academic success in engineering, were first-language English speakers. Thus, as the theory and research had suggested, a disciplinary-specific approach would potentially have the greatest impact in supporting undergraduates in engineering.

As early as 2012–2013, we had pilot tested two line graphs to replace the generic DELNA graphs. The new graphs illustrated changes in velocity over time in an acceleration test of a new vehicle. However, the graphs proved to be too difficult to interpret, given that they appeared on a writing assessment without any supporting context. Convinced that it was important to provide a writing task that better represented writing in engineering, in 2013 we also piloted and then administered a writing task that was embedded in the first lecture of a mandatory course, which all entering engineering students are required to take regardless of their future discipline (e.g., mechanical, aerospace, electrical engineering). During the first lecture, the professor introduced the topic, explained its importance, showed a video that extended the students' understanding of the topic, and announced that in the following class, the students would be asked to write about the topic by explaining differences in graphs which illustrated projected versus actual performance. Students were invited to review the topic on YouTube and given additional links for readings, should they choose to access them.

In 2014–2015, using the same embedded approach, we again administered the engineering graph task to 1014 students (99 % of the cohort). As in 2013, students wrote their responses to the diagnostic assessment in the second class of their required engineering course. The writing samples were far more credible and informative than had been the case with the generic task, which was unrelated to and unsupported by their academic work within the engineering program. The information produced by the hybrid rubric provided more useful information for peer mentors. Other diagnostic tasks were added to the assessment including a reading from the first-year chemistry textbook in engineering, with a set of multiple choice questions to assess reading comprehension, and a series of mathematics problems which represented foundational mathematics concepts required for first-year.

Thus, the initial generic approach evolved into a diagnostic procedure that was embedded within the university discipline of engineering and operationalized as an academic literacy construct as opposed to a language proficiency construct. The embedded approach is consistent with the theory that informed the study. As Figs. 3.1 and 3.2 illustrate, activities are situated within communities characterized by internal rules and a division of labour.

In the literature on post-entry diagnostic assessment, an embedded approach was first implemented at the University of Sydney in Australia as Measuring the Academic Skills of University Students (MASUS) (Bonanno and Jones 2007; Read 2015). Like the diagnostic procedure that is the focus of the present study, MASUS:

- operationalizes an academic literacy construct,
- draws on materials and tasks that are representative of a discipline,
- is integrated with the teaching of courses within the discipline, and
- is delivered within a specific academic program.

In December 2014, data from field notes collected by peer mentors indicated a dramatic increase in the number of students who were using the Centre (see details below). In large part, the establishment of a permanent support Centre, staffed with upper-year students in both engineering and in writing/language studies filled a gap in disciplinary support that had been evident for some time. The change in the mediational tools in the support activity system (i.e., to a permanent Centre), embedded within the context of the first-year required course in engineering, has also had an important impact on student retention and engagement.

5.3.2 Evidence Supporting a Permanent Place for Engineering Support: The Elsie MacGill Centre

In the context of voluntary uptake (Freadman 2002), where the decision to seek support is left entirely to the student, one of the greatest challenges was *reaching* students at-risk. As McLeod (2012) notes, such students are at times fearful, unsure, unaware, or unwilling to approach a Centre for help – particularly at the beginning of their undergraduate program. Only those students with exceptional social networking skills are likely to drop-in to a support Centre in the first weeks of a new year. These students manage a context adeptly (Artemeva 2008; Artemeva and Fox 2010) so that support works to their advantage (like the at-risk student who was the focus of McLeod's research).

From 2010 to 2012, the Centre providing support was open during the first 2 months of the Fall term (September–October) and was located in a number of different sites (wherever space was available). When the Centre closed for the year, interviews with engineering professors and TAs, instructors in engineering communications courses, and upper-year students who had worked in the Centre suggested the need for pedagogical support was on-going. Of particular note was the comment of one TA in a first-year engineering course who recounted an experience with a student who was failing. She noted, "I had no place to send him. He had no place to go." This sentiment was echoed by one of the engineering communications instructors who, looking back over the previous term, reported that one of her students had simply needed on-going support to meet the demands of the course. However, both she and her TA lamented their inability to devote more time to this student: "He was so bright. I could see him getting it. But there was always a line of other students outside my office door who also needed to meet with me. I just couldn't give him enough extra time to make a difference." Again, the type of support the student needed was exactly that which had been provided by the Centre. It was embedded in the context of engineering courses, providing on-going relevant feedback on engineering content, writing, and language.

As a result of evidence presented to administrators that the diagnostic assessment procedure and concomitant pedagogical support were having a positive impact, in 2013 a permanent space in the main engineering building was designated and named

the Elsie MacGill Centre[2] (by popular vote of students in the engineering program). It was staffed for the academic year by 11 peer mentors, 8 upper-year students from engineering and 3 from language/writing studies. In addition, funding was made available for on-going research to monitor the impact of the assessment procedure and pedagogical support.

From an Activity Theory perspective, the engagement of engineering students in naming, guiding, and increasingly using the Centre is important in understanding the evolution of the initial activities (see Figs. 3.1 and 3.2). As students increasingly draw on the interventions provided by mentors within the Centre (Fig. 3.2), motives of the two activities become more aligned and coherent; the potential for the novice's participation in the community of undergraduate engineering is increased because motives are less likely to conflict (or at least will be better understood by both mentors and students). As motives driving the activities of mentors and students increasingly align, the potential for positive impact on a student's experience, retention and academic success is also increased. Evidence of increasingly positive impact was gathered from a number of sources.

From September to December 2014, the peer mentors with an engineering background recorded approximately 135 mentoring sessions (often with one student, but also with pairs or small groups). However, the engineering peer mentors did not document whether students seeking support had been identified as at-risk.

During the same 3 month period, three students in the at-risk group made repeated visits (according to the log maintained by the writing/language studies peer mentors). However, not only at-risk students were checking in at the Centre and asking for help. There were 46 other students who used the pedagogical support provided by the writing/language studies peer mentors in the Elsie Centre (as it is now popularly called). In total, approximately 184 first-year students (19 % of cohort) sought pedagogical support in the first 3 months of the 2014–2015 academic year, and the number has continued to grow. Peer mentors reported that there were so many students seeking advice that twice during the first semester they had to turn some students away.

Increasingly, second-year students were also seeking help from the Elsie Centre. The majority were English as a Second Language (ESL) students who were struggling with challenges posed by a required engineering communications course. It was agreed, following recommendations of the engineering communications course instructors, that peer mentors would work with these students as well as all first-year students in the required (core) engineering course. In January 2015, one of the engineering communications instructors, with the support of the Elsie Centre, began awarding 1 % of a student's mark in the communications course for a visit and consultation at the Elsie Centre.

[2] Elsie MacGill was the first woman to receive an Electrical Engineering degree in Canada and the first woman aircraft designer in the world. She may be best known for her design of the Hawker Hurricane fighter airplanes during World War II. Many credit these small and flexible airplanes for the success of the Allies in the Battle of Britain. Students within the engineering program voted to name the Centre after Elsie MacGill.

Consistent with the theoretical framework informing this research, situating the diagnostic assessment procedure within a required engineering course has made a meaningful difference in students' voluntary uptake (Freadman 2002) of pedagogical support. As discussed above, in marketing the diagnostic assessment procedure to these engineering students, it was critical to work towards student ownership. Findings suggest that the students' increased ownership is leading to an important change in how students view and participate in the activity of the Centre. The engineering students in the 2014–2015 cohort seem to view the 'Elsie Centre' as an integral part of their activity system. As one student, who had just finished working on a lab report with a writing/language studies mentor, commented: "That was awesome. I'm getting to meet so many other students here, and my grades are getting better. When I just don't get it, or just can't do it, or I feel too stressed out by all the work…well this place and these people have really made a difference for me."

The Elsie Centre mentors have also begun to offer workshops for engineering students on topics and issues that are challenging, drawing on the personal accounts of the students with whom they have worked. The mentors have also undertaken a survey within the Centre to better understand what is working with which students and why. The survey grew out of the mentors' desire to elicit more student feedback and examine how mentors might improve the quality and impact of their pedagogical support. The activity system of the diagnostic assessment procedure (Fig. 3.1) is evolving over time, informed by systematic research, self-assessment, and the mentors' developing motive to be more effective in supporting more students. In other words, a more advanced activity system is emerging (Engeström 1987) which allows for further alignment in the activity systems of the mentors and the first-year students (Figs. 3.1 and 3.2).

6 Conclusion

Although the final figures for the 2014–2015 academic year are not yet available, there is every indication that the two changes made to the diagnostic assessment procedure, namely embedding the assessment in the content and context of a first-year engineering course, and setting up a permanent support Centre named and *owned* by these engineering students, have greatly increased its impact. Students are more likely to see the relevance and usefulness of diagnostic feedback and pedagogical support when it relates directly to their performance in a required engineering course. The Centre is open to *all* first-year students, and students of all abilities are using it. As a result, the Centre does not suffer from the stigma of a mandatory (e.g., remedial) approach. Increased engagement and participation by students is evidence of a growing interconnectedness that is shaping students' identities as members of the undergraduate engineering community. Lave and Wenger (1991) and Artemeva (2011) discuss the development of a *knowledgeably skilled identity* as an outcome of a novice's learning to engage with and act with confidence within a community. The development of this new academic identity (i.e., functioning

effectively as a student in engineering) contributes to the novice student's ability to act, increases the student's potential to learn (Artemeva 2011), and enhances their first-year experience. However, because use of the Centre remains voluntary, it remains an open question whether the Centre is reaching a sufficient number of students at-risk and is a focus of on-going research.

One of the most important outcomes of the two changes to the diagnostic assessment procedure was underestimated in its initial design. The Elsie Centre is facilitating the development of social connections within the engineering program, as students new to undergraduate engineering increase their sense of connection and community through interaction with peer mentors and their classmates in this newly created learning space. The findings from this study are consistent with the Vygotskian (1987) notions of knowledge and learning as situated and social. Empirical research on engagement identifies both academic and social considerations as key variables in predicting success in university study (e.g., Fox et al. 2014; Scanlon et al. 2007). Indeed, social connections that are fostered by interactions in the Elsie MacGill Centre may often be as important as academic connections in terms of enhanced first-year experience, retention, and levels of academic success. In the coming years, we will evaluate this relationship further with regard to the results of the tracking study, which will report on retention and academic success for new cohorts of entering undergraduate engineering students who have participated in the diagnostic assessment procedure described in this chapter.

References

Alderson, J. C. (2007). The challenges of (diagnostic) testing: Do we know what we are measuring? In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 21–39). Ottawa: University of Ottawa Press.

Anderson, T. (2015). Seeking internationalization: The state of Canadian higher education. *Canadian Journal of Higher Education, 45*(4), 166–187.

Artemeva, N. (2006). Approaches to learning genres: A bibliographical essay. In N. Artemeva & A. Freedman (Eds.), *Rhetorical genre studies and beyond* (pp. 9–99). Winnipeg: Inkshed Publications.

Artemeva, N. (2008). Toward a unified theory of genre learning. *Journal of Business and Technical Communication, 22*(2), 160–185.

Artemeva, N. (2011). "An engrained part of my career": The formation of a knowledge worker in the dual space of engineering knowledge and rhetorical process. In D. Starke-Meyerring, A. Paré, N. Artemeva, M. Horne, & L. Yousoubova (Eds.), *Writing in knowledge societies* (pp. 321–350). Fort Collins: The WAC Clearinghouse and Parlor Press. Available at http://wac.colostate.edu/books/winks/

Artemeva, N., & Fox, J. (2010). Awareness vs. production: Probing students' antecedent genre knowledge. *Journal of Business and Technical Communication, 24*(4), 476–515.

Artemeva, N., & Fox, J. (2014). The formation of a professional communicator: A sociorhetorical approach. In V. Bhatia & S. Bremner (Eds.), *The Routledge handbook of language and professional communication* (pp. 461–485). London: Routledge.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bonanno, H., & Jones, J. (2007). *The MASUS procedure: Measuring the academic skills of University students. A resource document.* Sydney: Learning Centre, University of Sydney. Retrieved from http://sydney.edu.au/stuserv/documents/learning_centre/MASUS.pdf

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 32–42.

Browne, S., & Doyle, H. (2010). *Discovering the benefits of a first year experience program for under-represented students: A preliminary assessment of Lakehead University's Gateway Program*. Toronto: Higher Education Quality Council of Ontario.

Cai, H. (2015). *Producing actionable feedback in EFL diagnostic assessment*. Paper delivered at the Language Testing Research Colloquium (LTRC), Toronto, 18 Mar 2015.

Cheng, L., & Fox, J. (2008). Towards a better understanding of academic acculturation: Second language students in Canadian universities. *Canadian Modern Language Review, 65*(2), 307–333.

Creswell, J. W. (2015). *A concise introduction to mixed methods research*. Los Angeles: Sage.

Elder, C., & von Randow, J. (2008). Exploring the utility of a web-based English language screening tool. *Language Assessment Quarterly, 5*(3), 173–194.

Engeström, Y. (1987). *Learning by expanding: An activity-theoretical approach to developmental research*. Helsinki: Orienta-Konsultit Oy.

Engeström, Y., & Miettinen, R. (1999). Introduction. In Y. Engeström, R. Miettinen, & R. I. Punamäki (Eds.), *Perspectives on activity theory* (pp. 1–16). Cambridge: Cambridge University Press.

Engeström, Y., Miettinen, R., & Punamäki, R. I. (Eds.). (1999). *Perspectives on activity theory*. Cambridge: Cambridge University Press.

Fox, J. (2009). Moderating top-down policy impact and supporting EAP curricular renewal: Exploring the potential of diagnostic assessment. *Journal of English for Academic Purposes, 8*(1), 26–42.

Fox, J. (2015). Editorial, Trends and issues in language assessment in Canada: A consideration of context. Special issue on language assessment in Canada. *Language Assessment Quarterly, 12*(1), 1–9.

Fox, J., & Artemeva, N. (2011). *Raters as stakeholders: Uptake in the context of diagnostic assessment*. Paper presented at the Language Testing Research Colloquium (LTRC), University of Michigan, Ann Arbor.

Fox, J., & Haggerty, J. (2014). *Mitigating risk in first-year engineering: Post-admission diagnostic assessment in a Canadian university*. Paper presented at the American Association of Applied Linguistics (AAAL) Conference, Portland.

Fox, J., Cheng, L., & Zumbo, B. (2014). Do they make a difference? The impact of english language programs on second language (L2) students in Canadian universities. *TESOL Quarterly, 48*(1), 57–85. doi:10.1002/tesq.

Fox, J., von Randow, J., & Volkov, A. (2016). Identifying students-at-risk through post-entry diagnostic assessment: An Australasian approach takes root in a Canadian university. In V. Aryadoust & J. Fox (Eds.), *Trends in language assessment research and practice: The view from the Middle East and the Pacific Rim* (pp. 266–285). Newcastle upon Tyne: Cambridge Scholars Press.

Freadman, A. (2002). Uptake. In R. Coe, L. Lingard, & T. Teslenko (Eds.), *The rhetoric and ideology of genre* (pp. 39–53). Cresskill: Hampton Press.

Gee, J. P. (2011a). *An introduction to discourse analysis: Theory and method* (3rd ed.). London: Routledge.

Gee, J. P. (2011b). *How to do discourse analysis: A toolkit*. London: Routledge.

Huhta, A. (2008). Diagnostic and formative assessment. In B. Spolsky & F. M. Hult (Eds.), *The handbook of educational linguistics* (pp. 469–482). Malden: Blackwell.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.

Leont'ev, A. N. (1981). The problem of activity in psychology. In J. V. Wertsch (Ed.), *The concept of activity in Soviet psychology* (pp. 37–71). Armonk: Sharp.

McLeod, M. (2012). *Looking for an ounce of prevention: The potential for diagnostic assessment in academic acculturation*. Unpublished M.A. thesis. Carleton University, Ottawa.

Office of Institutional Research and Planning. (2014). *Retention and Graduation of Undergraduates for "B. Engineering" – 1998 to 2012*. Ottawa: Carleton University.

Prior, P. (1994). *Writing/disciplinarity: A sociohistoric account of literate activity in the academy*. Mahwah: Lawrence Erlbaum.

Read, J. (2008). Identifying academic language needs through diagnostic assessment. *Journal of English for Academic Purposes, 7*(2), 180–190.

Read, J. (2012). *Issues in post-entry language assessment in English-medium universities*. Revised version of a plenary address given at the Inaugural Conference of the Association for Language Testing and Assessment of Australia and New Zealand (ALTAANZ). Australia, University of Sydney. 10 Nov 2012.

Read, J. (2015). *Assessing English proficiency for university study*. Gordonsville: Palgrave.

Scanlon, D. L., Rowling, L., & Weber, Z. (2007). ˜You don't have like an identity: you are just lost in a crowd": Forming a student identity in the first-year transition to university. *Journal of Youth Studies, 10*(2), 223–241.

Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition* (2nd ed.). Chicago: University of Chicago Press.

Vygotsky, L. S. (1987). Thinking and speech (N. Minick, Trans.). In R. W. Rieber & A. S. Carton (eds.), *The collected works of L. S. Vygotsky: Vol. 1. Problems of general psychology* (pp. 39–285). New York: Plenum Press. (Original work published 1934).

Chapter 4
The Consequential Validity of a Post-Entry Language Assessment in Hong Kong

Edward Li

Abstract The launch of the 3+3+4 education reform in Hong Kong has posed challenges to as well as created opportunities for tertiary institutions. It has invariably led to reviews of the effectiveness of their existing English language curricula and discussions among language practitioners in the tertiary sector as to what kind of English curriculum and assessment would serve the needs and interest of the new breed of senior secondary school graduates, who have had only six years to study English in the new education system as compared with seven years in the old system. This chapter reports on the pedagogical and assessment strategies adopted by the Hong Kong University of Science and Technology (HKUST) to embrace these challenges, and the findings of a pilot study conducted to investigate the consequential validity of a post-entry language assessment used at HKUST. Consequential validity is often associated with test washback. In Messick's expanded notion of test validity (Messick 1989), the evidential and consequential bases of test score interpretation and test score use are considered as crucial components of validity. It covers not just elements of test use, but also the impact of testing on students and teachers, the interpretation of test scores by stakeholders, and the unintentional effects of the test. This chapter reports the findings of the pilot study and discusses their implications for the use of PELAs.

Keywords Consequential validity • Post-entry language assessment • Hong Kong universities • Performance-based assessment • Criterion-referenced testing • Test-taker perceptions • Test washback • Assessment-curriculum integration

E. Li (✉)
Center for Language Education, The Hong Kong University of Science and Technology, Hong Kong, China
e-mail: lcedward@ust.hk

J. Read (ed.), *Post-admission Language Assessment of University Students*,
English Language Education 6, DOI 10.1007/978-3-319-39192-2_4

1 Introduction

This chapter reports on an exploratory study of some of the aspects of consequential validity of the English Language Proficiency Assessment (ELPA), a post-entry language assessment (PELA) developed by the Center for Language Education (CLE) at the Hong Kong University of Science and Technology (HKUST) for the first-year English Core curriculum. Under the policy of internationalization, the student intake at HKUST has become more diversified in the past decade or so. Although English language requirements are imposed for admission, over the years students have sought entrance via different pathways and with qualifications from various educational systems and regions. It would be a mistake to assume that these English standards and qualifications are largely equivalent, and that the students who have met any one of these requirements would have equally sufficient English to pursue their academic studies successfully. Similar concerns have been expressed by scholars in Australia and by the Australian Universities Quality Agency (Knoch and Elder 2013; Read 2015; AUQA 2009). In addition, the 3 + 3 + 4 educational reform in Hong Kong has created another layer of uncertainty about the university students' English proficiency profile at the entry point. The new educational system now sees secondary school students graduating a year earlier and universities offering 4-year undergraduate degrees in lieu of British-modeled 3-year degrees. Though central to the reform is a shift from a relatively intense subject specialization and examination-oriented learning culture to a more rounded general education curriculum with flexibility and choices for whole person development, the new academic structure also means that prospective university students have one year less in secondary school to develop their English proficiency before entering university.

Whether the students have achieved the proficiency threshold for effective academic studies in the medium of English has become a concern to the language practitioners in the Hong Kong tertiary sector. What could and should be done if admission requirements can only serve as a crude gatekeeping measure, while some incoming students may still fall through the cracks? For an English curriculum to work for all, how do we support those at risk and at the same time stretch the more able ones at the top? These challenges invariably necessitate a reconceptualization of the curriculum and assessment design to support students' English proficiency development after they enter university. In response to these challenges, some of the universities in Hong Kong choose to use homegrown PELAs as a pedagogical strategy to diagnose students' weaknesses or learning needs at entry and monitor their proficiency development at regular intervals of the undergraduate programmes. These PELAs are used in specific institutional contexts to suit specific institutional needs.

In Australia, similar concerns drive universities to seek more effective ways for students, non-native speaking students in particular, to develop their English language proficiency. The Australian Universities Quality Agency (AUQA) developed ten comprehensive, non-prescriptive Good Practice Principles for universities to follow with an aim to create a learning environment conducive to the development

of English language proficiency in university. Emphasis is given to how assessments can be used effectively to generate positive consequences on students' English language proficiency development. Out of these ten principles, five are quoted in this chapter for their direct relevance to the contexts of use of PELAs in Hong Kong in the face of the educational reform:

1. Universities are responsible for ensuring that their students are sufficiently competent in English to participate effectively in their university studies.
2. Resourcing for English language development is adequate to meet students' needs throughout their studies.
5. English language proficiency and communication skills are important graduate attributes for all students.
6. Development of English language proficiency is integrated with curriculum design, assessment practices and course delivery through a variety of methods.
7. Students' English language development needs are diagnosed early in their studies and addressed, with ongoing opportunities for self-assessment.

(AUQA 2009: 4)

2 The Teaching Context

The teaching context at HKUST resembles closely what these five Good Practice Principles recommend. The migration from the 3-year degree to the 4-year degree has led the University's senior management to reinstate the importance of English language development in the new 4-year undergraduate programme and the need for an English proficiency threshold for the progression to Year 2 onwards. Communication competence in English (and Chinese) is stated as one of the major learning outcomes of undergraduate programmes and as a graduate attribute for the 4-year degree. As indicated in Table 4.1, 12 out of 120 credits for an undergraduate programme are allocated to the development of English language ability and half of them to the first year English Core curriculum to help students build a solid foundation of proficiency and adequate academic literacy in English before they proceed to the senior years of studies. Six credits of study means six contact hours in the classroom with another six hours of out-of-class learning. The purpose of having a bottom-heavy English curriculum is to utilize the foundation year as much as possible to engage students intensively and actively in developing their English proficiency. At the same time, the Center for Language Education (CLE) also expands enormously the range of informal curricular language learning activities to cater for students' needs and interest. Students could choose to spend two full weeks in an

Table 4.1 The 12-credit English curriculum at HKUST

	Credits	Course length	English curriculum for the 4-year degree
Year 3/4	3	1 semester	Discipline-specific English courses
Year 2	3	1 semester	School-based English courses
Year 1	6	2 semesters	English core (with ELPA as the pre- and post-tests)

intensive on-campus immersion programme before the start of the first semester, participate in the academic listening workshops, enroll onto non-credit bearing short courses each targeting a specific language skill or aspect, join the regular theme-based conversation groups, or enjoy blockbusters in the movie nights. They could also seek one-on-one consultations with CLE advisors for more personalised help with their learning problems. In the 4-year degree programme, students are surrounded by various accessible learning opportunities in and out of the classroom.

In the English Core curriculum, ELPA serves both formative and summative assessment purposes. Like many other PELAs, ELPA is administered to students before the start of the first semester for early identification of the at-risk group who would need additional support for their English language learning. It also plays the role of a no-stakes pre-test to capture students' proficiency profile at the start point of their year-long English learning journey. The test results can be used as a reference point for students' self-reflection and choice of learning resources. At the end of the second semester of the English Core, ELPA is administered as a post-test to track students' proficiency gain throughout the year and to see if they have reached the proficiency threshold to proceed to Year 2. Unlike the pre-test, the ELPA post-test carries much higher stakes and the test results are counted towards the grades for the English Core. Because of its importance in the curriculum, care has been taken to incorporate assessment features in the design of ELPA that would maximize the possibility of positive washback on students and teachers.

Incorporating a test into a curriculum appropriately is never an easy task. Developing a homegrown PELA is even more daunting. Why reinvent the wheel? "If a test is regarded as important, if the stakes are high, preparation for it can come to dominate all teaching and learning activities. And if the test content and testing techniques are at variance with the objectives of the course, there is likely to be harmful backwash." (Hughes 2003:1) In the reconceptualization of the curriculum and assessment for the 4-year degree, ELPA as a PELA is designed in such a way as to:

- Capitalize on the flexibility and autonomy of an in-house PELA concerning what to assess and how to assess to serve the needs of the curriculum. ELPA is curriculum-led and curriculum-embedded;
- Help to constructively align teaching, learning and assessing. It should add value to consolidate the strengths of or even enhance the effectiveness of the English Core curriculum. The ultimate aim is that ELPA should generate positive washback on learning and teaching;
- Create a common language about standards or achievements between students and teachers. The senior secondary education examination results might not be appropriate as a reference to articulate expectations in the context of university study;
- Frame the assessment rubrics for the English Core course assessments; and
- Support teaching and learning by establishing a feedback loop. The test results should act as a guide to inform students about how to make better use of the learning resources at CLE and plan their study more effectively.

3 The Design of ELPA

ELPA is designed to assess the extent to which first-year students can cope with their academic studies in the medium of English. It assesses both the receptive (reading, listening and vocabulary) and productive (speaking and writing) language skills in contexts relevant to the academic studies. ELPA consists of two main test components. The written test assesses students' proficiency in reading, listening and writing, and their mastery of vocabulary knowledge. The speaking test is an 8-min face-to-face interview, assessing students' readiness to engage in a conversation on topics within the target language domain meaningfully and fluently (Table 4.2).

Despite the pre-test results being used as an indication of possible areas for more work by students, ELPA is not a diagnostic test by design. Typical diagnostic tests have the following characteristics (Alderson 2005: 10–12):

- They are more likely to focus on weaknesses than strengths. As a result, diagnostic tests may not be able to generate a comprehensive proficiency profile of students to inform decisions.
- They are likely to be less authentic than proficiency tests. Students would be required to complete tasks different to what they are expected to do in real life. Consequently, cooperation of students in taking the test may not be guaranteed.
- They are more likely to be discrete-point than integrative. This means diagnostic tests would be likely to focus more on specific elements at local levels than on global abilities. For the assessments of productive skills (speaking and writing) – the vital components of first-year progress – the discrete-point testing approach may not be most appropriate.

ELPA adopts mainly the performance-based test design and direct-testing approach, situating test items in academic contexts and using authentic materials whenever possible. What students are required to do in ELPA resembles closely the tasks they do in the real-life target domain. Assessments of the productive skills adhere to the direct testing approach. The prompts are generically academic or everyday social topics which an educated person is expected to be able to talk or write about. The input materials for the reading and listening sections are taken from authentic sources such as newspapers, magazines and journal articles, lectures, and seminars, with no or very slight doctoring to avoid content bias or minimize the

Table 4.2 The ELPA test

	Description	Format	Duration
Reading	Reading comprehension and text reconstruction	Multiple-choice	40 min
Listening	Social conversations; consultations; seminars; lectures	Multiple-choice	30 min
Vocabulary	Lexical recognition and recall	MC and gap fill	40 min
Writing	Expository essay of 300 words	Essay	40 min
Speaking	Interview and recommending solutions to a problem	Interview	8 min

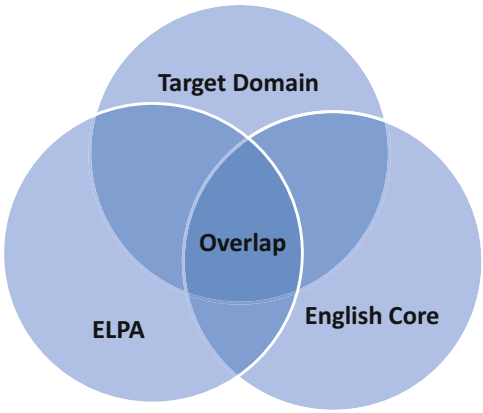


Fig. 4.1 Overlap of constructs between ELPA, English core and the target language domain

need for prior knowledge. The test domain for the assessment of vocabulary knowledge is defined by theAcademic Word List and the benchmark word frequency levels for the junior and senior secondary education in Hong Kong.

As pointed out by Messick (1996), possessing qualities of directness and authenticity is necessary but not sufficient for a test to bring beneficial consequences. The test should also be content and construct valid. In the case of ELPA, the construct andcontent validity is addressed by full integration of the test and the curriculum, using the list of most desirable constructs for development not only as the specifications for test development but also the curriculum blueprint for materials writing and articulation of learning outcomes. This is to ensure as high a degree as possible of critical alignment of the assessment, teaching and learning components of the curriculum foundation. Although the curriculum has a broader spectrum of contents and objectives than those of the ELPA test specifications, all the major constructs intended for assessment in ELPA are significantly overlapped in the curriculum. The constructs are assessed in ELPA as snapshots and in course assessments as achievements over time for formative and summative feedback. Overlap just between the test and the course is not automatically equivalent to content validity. The latter also involves the relationships of curriculum and assessment with the target language domain, as suggested by Green (2007) for positive washback (Fig. 4.1)

ELPA also adopts the criterion-referenced testing approach. The ELPA rating scale is made up of seven levels, with Level 4 set as the threshold level for the English Core curriculum. The performance descriptors for each level are written as can-do statements with typical areas of weaknesses identified for further improvement. The performance descriptors are also used in an adapted form as the assessment rubrics for the course assessments of the English Core. This is to establish a close link between ELPA and the coursework. Both the ELPA test results and the feedback on course assignments are presented to students in the form of ELPA levels attained and corresponding performance descriptors. After the pre-test results are released, there will be individual consultations between students and their class

teacher to negotiate a study plan in the form of a learning contract to improve on areas where further language work is warranted. Students have complete freedom to choose whether they prefer working on their own with the materials they find for themselves, talk to the language advisors in CLE's Language Commons for further guidance, or take some of the short non-credit bearing courses or on-going activities to address their language needs. Support for students on their language needs is in plentiful supply at HKUST. The real issue is not availability of such language support for students, but more with its utility.

4 A Framework for Test Validation

Messick (1989, 1996) argues the need in test validation studies to ensure that the social consequences of test use and interpretation support the intended purposes of testing and are consistent with the other social values. He believes that the intended consequences of a test should be made explicit at the test design stage and evidence then be gathered to determine whether the actual test effects correspond to what was intended. Regarding what evidence for validating the consequences of test use should be collected and how the evidence should be collected, Messick's theoretical model of validity does not give much practical advice for practitioners. His all-inclusive framework for validating the consequences of score interpretation and use requires evidence on all the six contributing facets of validity. There is no indication as to the operationalization and prioritization of the various aspects of validity to justify the consequences of test use.

On the other hand, Weir (2005) offers a more accessible validation framework as an attempt to define and operationalize the construct. He proposes three possible areas to examine Messick's consequential validity: differential validity – or what he calls avoidance of test bias in his later writings (Khalifa and Weir 2009; Shaw and Weir 2007); washback on individuals in the classroom/workplace; and impact on institutions and society. As can be seen, test fairness is a key component of this framework. Weir argues that one of the basic requirements for a test to create beneficial effects on stakeholders is that the test has to be fair to the students who take it, regardless of their backgrounds or personal attributes. If the intended construct for assessment is under-represented in the test or the test contains a significant amount of construct-irrelevant components, the performance of different groups of students will likely be affected differentially and the basis for score interpretations and use would not be meaningful or appropriate (American Educational Research Association et al. 1999). This echoes Messick's argument for authenticity and directness as validity requirements by means of minimal construct under-representation and minimal construct-irrelevant items in a test.

Weir's validation framework also makes a distinction between washback and impact by relating the former to the influences the test might have on teachers and teaching, students and their learning (Alderson and Wall 1993), and defining the latter as the influences on the individuals, policies and practices in a broader sense

Table 4.3 An expanded framework for consequential validation of PELAs

Aspects of consequential validity	Possible areas for investigation
Institutional context for test use	English language policy for effective study for all students
	Diagnosis of English language development needs
	Opportunities for students' self-assessment
	Integration of English language proficiency development with curriculum design, assessment practices and course delivery
	Support for students for language development needs
	Support for teachers
Intended validity (e.g. test design, test content and logistical strategies)	Test design (e.g. the use of assessment approaches)
	Intended construct validity (i.e. the most desirable constructs in the target domain/curriculum to test)
	Intended content validity (e.g. bias, sampling, overlap between test and target/curriculum domains)
	Test formats (e.g. use of a range of formats)
	Authenticity of materials and tasks
	Stakeholders' knowledge of and involvement in the test
Washback – perceptions of stakeholders	Test difficulty
	Test importance
	Attainability of test success
	Content validity
	Test fairness
	Construct validity
Washback – actions taken by stakeholders	Influence on teaching (e.g. attitude, content, method and outcomes)
	Influence on learning (e.g. attitude, content, method and outcomes)
Effect on individuals within society	Impact on curriculum, policies and the institution

(Wall 1997; Hamp-Lyons 1997; Bachman and Palmer 1996), though washback and impact are not two entirely unrelated constructs. As Saville (2009:25) points out, "the impact: washback distinction is useful for conceptualizing the notion of impact, but it does not imply that the levels are unconnected." Building on Messick and Weir's ideas above, an expanded validation framework for evaluating the consequential validity of PELAs is proposed in Table 4.3, with avoidance of test bias subsumed under test content and washback divided into two related categories: perceptions of the test and actions taken as a result of test use (Bailey 1996).

5 Consequential Validity of ELPA

This section reports on the first phase of an exploratory longitudinal study of the consequential validity of ELPA. The findings are presented along the dimensions of the expanded validation framework.

5.1 Context for Test Use and Intended Validity

As stated in Sect. 3, the institutional context in which ELPA is used mirrors the recommendations of the Good Practice Principles by AUQA (2009). One aspect that might seem to have fallen short of Good Practice Principle No. 7 is that students only take ELPA twice, first before the start of Year 1 and then at the end of Year 1. It might be considered as not providing the on-going opportunities for students' self-assessment. The situation at HKUST is that the reflection on one's language development needs does not happen through re-sitting ELPA multiple times for the purposes of self-assessment. This is different from the use of DELTA, for example, a PELA used in some sister institutions in Hong Kong, which is intended to track students' proficiency development at multiple points throughout their degree study (see Urmston et al. this volume). In the case of ELPA, where assessment is fully integrated with the curriculum, opportunities for self-reflection about language development needs and further language work are provided in the context of the course, particularly when the class teacher gives feedback to students on their course assignment using the ELPA assessment framework.

The intended validity was examined by the ELPA test development team, the English Core curriculum team and an external consultant in the form of a reflection exercise. The intended or a priori validity evidence (Weir 2005) gathered before the test event can be used to evaluate whether the test design has the potential for beneficial consequences. The group generally felt that ELPA follows the sound principles of performance-based, criterion-referenced, and direct-testing approaches. Given the limitations of PELAs, the challenge for the ELPA team is how to ensure that the test assesses the most relevant and desirable constructs and can "sample them widely and unpredictably" (Hughes 2003: 54) in the target language domain without making the test too long or having students take the test multiple times before inferences can be made. There is always tension between validity, reliability and practicality. Our response to this challenge is that ELPA and the curriculum complement each other in plotting the different facets of students' proficiency development – as test-takers vs learners; in test-based snapshots vs extended process assessments; and in examination conditions, with no learning resources vs in resource-rich classroom assessment situations where students can be engaged in collaborative assessments with peers.

5.2 Perceptions of Students

Students' perceptions of test validity are one important source of evidence to support a test use, and arguably even more influential than the statistical validity evidence in explaining their willingness to take part in the testing event and shaping their learning behaviors afterwards. More than 1,400 incoming students participated in a survey immediately after they took ELPA as a pre-test after admission in

summer 2014. The questionnaire covered aspects including test administration and delivery, affective factors such as anxiety and preferences for test formats, and the test quality. Since students had not yet started the English Core courses, questions related to content validity or the degree of overlap between the test, the course and the English ability required in the academic courses were not included in this survey. Twenty students were randomly selected for two rounds of focus discussions in October 2014, after two months of degree studies, to give more detail about the effect of ELPA on their perceptions or attitudes and the actions they took as a result of taking the test.

Reported in this section are the findings on perceptions of test difficulty, test importance, test fairness/bias, and validity. To reveal the possible differences between students from different schools at HKUST, the data were subjected to one-way ANOVAs. The ratings of each related item were analysed as dependent measures and "SCHOOL" (Business, Engineering, Science and Humanities) was a between-subject factor. Post-hoc Tukey HSD tests were conducted with ratings between different schools when a significant main effect of SCHOOL was found. The significance of the post-hoc tests was corrected by the Holm-Bonferroni method.

Overall, students found the ELPA test somewhat between '3=appropriate' and '2=difficult' (mean = 2.43; SD = 0.766) in terms of test difficulty (Table 4.4). One-way ANOVAs were conducted with ratings for each sub-test. The results showed that there were significant main effects of SCHOOL ($p < 0.05$) in most cases, with the only exception in Reading ($p = 0.09$). Post-hoc tests indicated that significant differences in SCHOOL mainly existed between Business and Engineering students ($p < 0.01$), with Engineering students finding the test more difficult. Writing and Speaking were most positively rated as 'appropriate' but Vocabulary Part 2 as significantly more difficult than other parts.

To check if Writing and Speaking were statistically different from other parts, the data of 1,049 students who rated all six parts were subjected to repeated-measures ANOVA. The ratings of each part were analysed as dependent measures, with "TEST" (Reading, Listening, Vocabulary Part 1, Vocabulary Part 2, Writing and Speaking) as a within-subject factor. Results showed that there was a significant main effect of TEST ($F(4.57, 4786.85) = 351.04$, $p < 0.001$. Greenhouse-Geisser correction was used for sphericity). Post-hoc paired t-tests were conducted to compare the ratings for Writing and Speaking to those for the other parts (significance was corrected by the Holm-Bonferroni method). Results of the post-hoc tests showed that Writing and Speaking were rated higher, and closer to "appropriate", ($p < 0.001$) than other parts, but no difference was found between Writing and Speaking.

This pattern can be attributed to students' familiarity with and preference for test formats. In the subsequent focus group discussions, students commented that they were familiar and comfortable with the essay writing and interview test formats. As they said, the test formats for Writing and Speaking were what they 'expected'. Vocabulary Part 2, on the other hand, consisted of 30 sentences with the target word in each of them gapped. Students had to understand the given context and then fill in the most appropriate word to complete the meaning of the sentence. Not only did students say they were not familiar with this test format but recalling the target word

Table 4.4 Students' perception of test difficulty

Question: Is the ELPA test easy? (5 = very easy; 3 = appropriate; 1 = very difficult)															
	Overall			Business			Engineering			Science			Humanities		
	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N
ELPA test	2.43	0.77	1462	2.52	0.79	474	2.34	0.76	572	2.43	0.75	378	2.63	0.63	38
Reading	2.54	0.85	1469	2.51	0.83	474	2.50	0.87	573	2.63	0.87	383	2.64	0.74	39
Listening	2.30	0.90	1470	2.37	0.89	475	2.24	0.91	573	2.27	0.91	383	2.62	0.78	39
Vocab Pt. 1	2.75	0.99	1469	2.93	0.96	476	2.63	1.00	572	2.68	1.00	383	2.89	0.86	38
Vocab Pt. 2	2.12	0.95	1469	2.29	0.94	475	1.99	0.96	573	2.09	0.93	382	2.36	0.81	39
Writing	3.05	0.88	1469	3.14	0.88	475	2.96	0.89	574	3.06	0.84	381	3.21	0.89	39
Speaking	3.07	0.87	1061	3.10	0.85	365	3.02	0.91	381	3.07	0.82	289	3.54	0.95	26

from their mental lexicon to match the given context was stress-provoking. The less common the target words, the more stress they would cause for students. Vocabulary Part 1 seemed more 'manageable' as choices were provided to match with the meaning of the target word. At least the students thought the multiple choice format induced less stress.

The content of the writing and speaking parts of the test was considered as 'neutral' and 'impartial' (Table 4.5: written test – mean = 3.09, SD.668; speaking test – mean = 3.06, SD = 0.670). No significant main effect of SCHOOL was found for either test (written test: ($F(3,1462) = 0.58$, $p = 0.63$); speaking test: ($F(3,1087) = 1.21$, $p = 0.31$)). These questionnaire findings were confirmed by nearly all the participants in the focus groups. The writing and speaking prompts they were given in these two parts were topics which they said concerned their everyday life and therefore they had views to express. They did not feel they were 'being tricked or trapped', or that the topics gave them any advantage or disadvantage in terms of prior knowledge. Choice of topic was mentioned as a potential area for improving the social acceptability of ELPA. For example, students understood that it might not be possible to have more than one topic or theme to choose from in a live speaking test, but they preferred a choice of topics in the writing paper if possible. This was because they were used to choices in the Hong Kong Diploma of Secondary Education (HKDSE) Examination which adopts the graded approach where candidates could choose to attempt the easier or the more challenging versions of some of the sections. They could also choose from a range of topics for the HKDSE Writing paper.

The perceived validity of ELPA was rated between '3=accurately measured my English ability' and '2=my performance was somewhat worse than my English ability' consistently across all papers and backgrounds (Table 4.6). No significant main effect of SCHOOL ($p > 0.05$) was found in any of the tests.

They generally felt satisfied that the test assessed their English proficiency to a reasonably good extent, though more than two-thirds of the participants in the focus groups said that they could have done better. Because of this 'can-be-better' mentality, they chose '2' rather than a '3' to this question.

Overall, the students thought the test was fair and it was important for them to get good results in ELPA (Table 4.7). The one-way ANOVA revealed a significant effect of SCHOOL ($F(3,1464) = 5.79$, $p = 0.001$). Post-hoc Tukey HSD tests indicated that Business students (mean = 4.38 ± 0.79) perceived getting good results in ELPA as more important than Science students did (mean = 4.23 ± 0.89, $p < 0.001$), among all comparisons. As both means are above 4, this can be interpreted as a difference in the degree of importance rather than in whether it was considered important or not.

Table 4.5 Students' perception of content bias

Question: Does the test content give you more advantage than other students? (5 = I have advantage over other students; 3 = I do not have any advantage or disadvantage; 1 = I have disadvantage over other students)															
	Overall			Business			Engineering			Science			Humanities		
	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N
Written test	3.09	0.67	1466	3.07	0.66	473	3.08	0.66	573	3.12	0.69	381	3.15	0.71	39
Speaking test	3.05	0.67	1091	3.06	0.69	368	3.02	0.68	401	3.06	0.62	294	3.25	0.80	28

Table 4.6 Students' perception of validity

Question: Did the test accurately measure your English ability? (5 = my performance was better than my ability; 3 = my performance accurately reflects my ability 1 = my performance was worse than my ability)														
	Overall			Business			Engineering			Science			Humanities		
	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N
ELPA test	2.54	0.68	1447	2.58	0.68	466	2.54	0.69	566	2.52	0.68	377	2.50	0.65	38
Reading	2.65	0.68	1456	2.63	0.69	470	2.67	0.67	568	2.67	0.68	379	2.59	0.55	39
Listening	2.50	0.76	1458	2.49	0.78	470	2.52	0.74	570	2.46	0.78	379	2.59	0.60	39
Vocab Pt. 1	2.77	0.66	1457	2.80	0.68	470	2.79	0.63	569	2.78	0.68	379	2.72	0.46	39
Vocab Pt. 2	2.60	0.71	1455	2.65	0.70	470	2.60	0.72	568	2.52	0.73	378	2.56	0.50	39
Writing	2.68	0.72	1455	2.71	0.72	468	2.67	0.72	570	2.69	0.73	378	2.64	0.58	39
Speaking	2.64	0.73	1095	2.68	0.72	374	2.65	0.73	401	2.56	0.74	292	2.61	0.57	28

Table 4.7 Students' perception of test importance

Question: How important is it for you to get good results in ELPA? (5 = very important; 3 = somewhat important; 1 = not important at all)															
	Overall			Business			Engineering			Science			Humanities		
	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N
ELPA test	4.26	0.87	1468	4.38	0.79	474	4.23	0.89	573	4.14	0.92	382	4.31	0.89	39

5.3 Washback on Students

In the focus group discussions, students expressed mixed feelings about the test. Three out of the 20 students in the discussions held rather negative views. The fact that they had already been admitted to the university meant that they felt they should have satisfied the University's English language requirement. While they accepted the need for the English Core for first year students, they did not quite understand why ELPA was necessary after admission. To them diagnosis was not meaningful because they said they had very good ideas about their strengths and weaknesses in English. When prompted to state what their weaknesses were, they made reference to their results in the English Language subject in the HKDSE examination. Their resentment seemed to be related to their perception of the policy's fairness. Interestingly, their ELPA pre-test results showed that they did not belong to one single ability group – one achieved relatively high scores in ELPA on all sections, one mid-range around the threshold level and one below the threshold. They said ELPA had no influence on them and would not do anything extra to boost their test performance.

Other focus group participants accepted ELPA as a natural, appropriate arrangement to help them identify their language needs for effective study in the medium of English. They agreed that the learning environment in the university was much more demanding on their English ability than that in secondary school. For example, they were required to read long, unabridged authentic business cases or journal articles and listen to lectures delivered by professors whose regional accents might not be familiar to them. They showed enhanced motivation to do well in the ELPA post-test not just for English Core but also for other academic courses, though to some this attitude might have been caused by increased anxiety and trepidation rather than appreciation of opportunities for personal development.

When asked whether they had done more to improve their English proficiency than before as a result of ELPA, the 17 student participants who had a more positive attitude towards ELPA said they had spent more time on English. However, the washback intensity of the pre-test was obviously not very strong. They said this was partly because the ELPA post-test was not imminent and partly because English Core already took up a substantial amount of their study time allocated to language development. They knew improvement of English proficiency would take a long time but unanimously confirmed that they would do more on the requirements of ELPA to get themselves prepared for the ELPA post-test. The washback intensity was expected to become stronger in the second semester when the ELPA post-test was closer in time.

Regarding what to learn and how to learn for ELPA, 15 student participants chose expanding their vocabulary size as their top priority. They reasoned that vocabulary was assessed not just in the vocabulary section, but also the writing and speaking sections of the test. In addition, they felt that Vocabulary Part 2 (Gap Fill) was the most difficult section of the test. Improving their lexical competence would be the most cost-efficient way to raise their performance in ELPA, though they

doubted whether they could get better scores in Vocabulary Part 2 because of its unpredictability. They also thought learning words was more manageable in terms of setting learning targets and reducing the target domain into discrete units for learning than the more complex receptive and productive language skills. They all started with, probably under their teachers' advice, the items on the Academic Word List, a much smaller finite set of words than other word frequency levels. They made use of online dictionaries for definitions and the concordance-based learning materials to consolidate their understanding of the words in relation to their usage and collocation. Two students who were from Mainland China said that they kept a logbook for vocabulary learning, which included not just the English words but also words and expressions of other languages such as Cantonese, the spoken language used by the local Hong Kong people.

Apart from vocabulary knowledge, eight participants chose writing as another major learning target for the first semester. Six focused on expanding the range of sentence structures they could use in an essay whereas four identified paragraph unity as their main area for further improvement. They said that these were criterial features of the ELPA performance descriptors for writing and were taught in class. They said they had a better idea of how to approach the learning targets – they followed the methodology used in the textbook for practice, for example, varying sentence patterns instead of repeating just a few well-mastered structures, writing clear topic sentences and controlling ideas for better reader orientation. However, they said that teachers' feedback on the practice was crucial for the sustainability of the learning habits.

5.4 Washback on Teachers

Six teachers agreed to join a focus group discussion to share their views on ELPA and how the test affected their perceptions and the way they taught in the classroom. Overall the six teachers had a very positive attitude towards ELPA and its role in the English Core curriculum. Nearly all teachers on the English Core teaching team are ELPA writing and speaking assessors. Many of them had been involved in item writing and moderation at some stage, and a few were designated to carry out more core duties in the assessor training, quality assurance, and test administration and operation. Their minor grievances seemed to be all related to the extra workload caused by ELPA, though these extra duties were all compensated with teaching relief. However, they all commented ELPA had helped to make the assessment component of the course more accurate and the assessment outcomes more credible. They felt that their involvement in different aspects of ELPA had helped them gain a better understanding of the assessment process. For example, they found the elaborate Rasch analysis-based rater anchoring plans for the speaking and writing assessments extremely tedious, yet necessary for quality assurance.

Despite a heavier assessment load for the assessors, they agreed that the whole procedure helped to minimize teacher variability in assessment across classes and

as a result they had more confidence in the assessment outcomes. Another aspect that they appreciated most was the integration of criterion-referenced ELPA performance descriptors into course assessment. They said ELPA performance descriptors evolved into a common language with explicit, mutually understood reference points among teachers and between teachers and students to discuss standards and achievements. This was very useful for them to formulate teaching targets and to help students to articulate needs and set milestones for learning. They felt that student extrinsic motivation was enhanced.

All the teacher participants agreed that ELPA was a high-stakes test in the English Core, so naturally they considered it their duty to help students to at least meet the threshold requirement by the end of the second semester. Regarding the washback on teachers as to what to teach and how to teach, the influence from ELPA was less clear. Two teachers pointed out that they would recommend that students be more exposed to authentic English and practice using English whenever and wherever possible, whether or not ELPA was an assessment component of the course. They believed that an increase in proficiency would naturally lead to better performance in ELPA, so what they did in the classroom was simply based on the principles and good practice of an effective EFL classroom. ELPA could be one of the more latent teaching goals, but certainly was not to be exploited as the means for teaching.

While the other four teacher participants agreed that this naturalistic approach was also what they followed in the classroom to help students to develop their *receptive* skills, they saw more commonality between the sections of the ELPA test on *productive* skills and vocabulary knowledge and the curriculum, so that could form a useful basis for teaching that served the purposes of both. For example, topic development and coherence is one of the assessment focuses of both the ELPA speaking and writing tests. What these teachers said they did was teach students strategies for stronger idea development which were emphasized in the ELPA performance descriptors, i.e., substantiation of arguments with examples and reasons, rebuttal, paragraph unity with clear topic sentences and controlling ideas. They argued that this approach would contribute to desirable performance in both the test and the course assessment. Asked if they would teach in the same way were ELPA not a component of the course, they admitted that they would do more or less the same but the intensity perhaps would be much weaker and their focus would probably be only on the tasks required in the course assessment rather than a wider range of tasks similar to those used in the ELPA test.

5.5 Impact

The first impact that ELPA has created is in the resources it demands for maintenance of the current scope of operation, i.e. from test development, test administration, management of a team of assessors, statistical analysis, quality assurance to score and result reporting. It is a resource-hungry assessment operation. Yet it is a

view shared by the development team, the curriculum team, and the Center senior management that the current assessment-curriculum model adopted is the best possible approach in our context to ensure a high degree of overlap between the test, the curriculum and the target language domain for positive washback. Colleagues in CLE would regard their various involvements in ELPA as opportunities for personal and professional development, from which the gains that they might obtain could benefit the development and delivery of other courses.

Because of the high stakes of ELPA in the English Core curriculum, a policy has been put in place to counteract the probabilistic nature of assessments – all borderline fail cases in ELPA and English Core are required to be closely scrutinized to ensure accuracy in the evidence and appropriateness in the decisions made. Those who fail to meet the threshold by the second semester can retake the course and the test in the summer. Classes for repeaters will be much smaller in size to provide more support and attention for the students, and they will have a range of options as to what types of out-of-class support they need and prefer. Evidence from the classroom about achievements and learning progress is also gathered to support any further decisions to be made.

6 Conclusion

Evaluation of the consequential validity of a PELA is a daunting task. Each of the facets of the consequential validity in the expanded validation framework (Table 4.3) poses its own challenges to the researcher. The notion of washback of a PELA is particularly messy because it has to be conceived as a comparative phenomenon so as to observe the influence a test has on the stakeholders and how it affects actions that they would not necessarily take (Alderson and Wall 1993; Messick 1996). The evidence collected in the current study seems to suggest that ELPA has generated positive consequences for the stakeholders. However, some of the observations are inconclusive. For example, we cannot be sure whether what teachers do in the classroom is a function of variables such as teacher training, teachers' preference or the influence from ELPA. We asked students and teachers hypothetical questions, 'Would you…if ELPA were not part of the curriculum?' Our practical constraint is that investigations could not be conducted according to a tight experimental design: we could not give ELPA to one group of students and not to another group within the same cohort for the sake of comparison. Studies of the consequential validity of PELAs have to be based on understandings of how a test functions within a specific institutional context. The findings reported in this chapter are the first phase of a longitudinal study. Triangulation through classroom observations, investigations into the students' learning patterns and preferences, comparison and analysis of students' performances in not just the test and the curriculum but also the target domain use would be useful evidence to demonstrate the beneficial consequences of test use.

References

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. New York: Continuum.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115–129.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC.

AUQA (Australian Universities Quality Agency). (2009). *Good practice principles for English language proficiency for international students in Australian universities*. Report to the Department of Education, Employment and Workplace Relations, Canberra.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing, 13*(3), 257–279.

Green, A. (2007). *IELTS washback in context: Preparation for academic writing in higher education* (Studies in Language Testing, Vol. 25). Cambridge: UCLES/Cambridge University Press.

Hamp-Lyons, L. (1997). Washback, impact and validity: Ethical concerns. *Language Testing, 14*(3), 295–303.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading* (Studies in Language Testing, Vol. 29). Cambridge: UCLES/Cambridge University Press.

Knoch, U., & Elder, C. (2013). A framework for validating post-entry language assessments (PELAs). *Papers in Language Testing and Assessment, 2*(2), 48–66.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: American Council on Education and Macmillan Publishing Company.

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

Read, J. (2015). Plenary speech: Issues in post-entry language assessment in English-medium universities. *Language Teaching, 48*(2), 217–234.

Saville, N. (2009). *Developing a model for investigating the impact of language assessment within educational contexts by a public examination provider*. Unpublished PhD thesis, University of Bedfordshire.

Shaw, S., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing* (Studies in Language Testing, Vol. 26). Cambridge: UCLES/Cambridge University Press.

Wall, D. (1997). Impact and washback in language testing. In C. Clapham & D. Corson (Eds.), *Language testing and assessment* (pp. 291–302). Amsterdam: Kluwer Academic Publishers.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. New York: Palgrave Macmillan.

Chapter 5
Can Diagnosing University Students' English Proficiency Facilitate Language Development?

Alan Urmston, Michelle Raquel, and Vahid Aryadoust

Abstract The effectiveness of a language test to meaningfully diagnose a learner's language proficiency remains in some doubt. Alderson (2005) claims that diagnostic tests are superficial because they do not inform learners what they need to do in order to develop; "they just identify strengths and weaknesses and their remediation" (p. 1). In other words, a test cannot claim to be diagnostic unless it facilitates language development in the learner. In response to the perceived need for a mechanism to both provide diagnostic information and specific language support, four Hong Kong universities have developed the Diagnostic English Language Tracking Assessment (DELTA), which could be said to be meaningfully diagnostic because it is both integrated into the English language learning curriculum and used in combination with follow-up learning resources to guide independent learning.

This chapter reports ongoing research into the effectiveness of DELTA to support students' efforts to improve their English language proficiency while at university. Specifically, the chapter reports on a study of students who have taken DELTA more than once and for whom it is possible to measure growth through the DELTA's use of Rasch modeling. The students were surveyed to determine the English language enhancement activities that they engaged in during the period in which the growth was observed, including their use of the report provided to them after taking the DELTA.

Keywords Diagnosis • Tracking • Rasch modeling • Development • Enhancement • Intervention

A. Urmston (✉)
English Language Centre, Hong Kong Polytechnic University, Hong Kong, China
e-mail: alan.urmston@polyu.edu.hk

M. Raquel
Centre for Applied English Studies, University of Hong Kong, Hong Kong, China
e-mail: michelle.raquel@hku.hk

V. Aryadoust
National Institute of Education, Nanyang Technological University, Singapore, Republic of Singapore
e-mail: vahid.aryadoust@nie.edu.sg

J. Read (ed.), *Post-admission Language Assessment of University Students*, English Language Education 6, DOI 10.1007/978-3-319-39192-2_5

1 Introduction

At Hong Kong universities, undergraduates are provided with English language enhancement which mainly focuses on English for Academic Purposes (EAP). However, it has long been recognised that students require not only EAP, but also what might be regarded as more general or non-specific English language enhancement to raise their proficiency to a level at which they can function within an English-medium environment (Evans and Green 2007). Universities in Hong Kong now receive an annual grant from the Universities Grants Committee for such provision in the form of taught credit and non-credit bearing courses, self-access learning facilities, and extra-curricular activities. The kind of provision varies from institution to institution though, as mentioned, EAP is the core, with students typically learning the skills of academic writing of different genres (e.g. problem-solution texts, discursive essays), referencing conventions, academic writing style and presentation skills (Evans and Morrison 2011). In some cases, such EAP courses are generic across disciplines, while in others they are discipline-specific. In addition, each university has self-access language learning facilities to provide support of these EAP courses and encourage students to improve on their own,[1] and extra-curricular activities such as clubs and societies through which they can use the language in less formal contexts.[2] Finally, a number of universities provide more individualised learning programmes, often involving mentoring from peer students or English teachers. One such programme, in operation at the Hong Kong Polytechnic University, the Excel@English Scheme (EES), uses diagnostic assessment (DELTA – see below) to inform individualised learning programmes that students can follow and tracks their progress through it.[3]

As numerous as the English language enhancement activities are the assessments that students need to take. These include course-embedded tests and assessments as well as assessments designed for a variety of other purposes. A list of the course-embedded tests and assessments in use at the Hong Kong Polytechnic University highlights the extent and variety of testing that goes on, particularly of writing (Table 5.1).

The majority of these assessments could be described as being summative in nature and are normally a judgment which encapsulates all the evidence up to a given point. Even though in some cases with writing assessments a process approach is adopted and students are given feedback on drafts before being assessed on the final version, ultimately the focus is on improving the product rather than developing skills, since courses are credit-bearing and grades need to be counted.

On the other hand, *formative assessment* is specifically intended to generate feedback on performance to improve and accelerate learning (Sadler 1998). Black and Wiliam (1998) carried out a substantial review of formative assessment and concluded that it is effective in promoting student learning across a wide range of

[1] For a detailed description of the self-access centres in universities in Hong Kong, see Miller and Gardner (2014).

[2] See, for example, http://elc.polyu.edu.hk/Clubs-Societies

[3] For more information on EES, see http://elc.polyu.edu.hk/EES/

Table 5.1 English language assessment tasks in use at the Hong Kong Polytechnic University

Assessments of writing	**Assessments of speaking**
Analysing texts of different genres	Debates
Compiling writing portfolios	Group discussions
Reflective writing	Job interviews
Writing blogs	Meetings
Writing book reports	Oral presentations
Writing business reports	
Writing case reports	**Other assessments**
Writing design project reports	Digital stories
Writing discursive essays	Grammar tests
Writing emails and letters	Listening tests
Writing expository essays	Online assessments
Writing feature articles	Reading tests
Writing persuasive essays	Vocabulary tests
Writing position-argument essays	
Writing problem-solution essays	
Writing project reports	
Writing promotional literature	
Writing proposals	
Writing reports with reference to sources	
Writing short reports	
Writing short stories	
Writing short texts	
Writing speeches	
Writing technical texts	
Writing workplace-oriented texts	

education settings (disciplinary areas, types of outcomes, levels). Formative assessment requires feedback which indicates the existence of a 'gap' between the actual level of the work being assessed and the required standard. It also requires an indication of how the work can be improved to reach the required standard (Taras 2005). Such assessment is critically important for student learning. Without formative feedback on what they do, students will have relatively little by which to chart their development (Yorke 2003). Another central argument is that, in higher education, formative assessment and feedback should be used to empower students as self-regulated learners. In Hong Kong, while there have been in place common summative assessments of university entrants'[4] and graduates'[5] English proficiency,

[4] The Hong Kong Diploma of Secondary Education (HKDSE) and previously the Hong Kong Advanced Level Use of English examinations.

[5] Between 2003 and 2014, the Hong Kong Government employed the International English Language Testing System (IELTS) under its Common English Proficiency Assessment Scheme,

what has been missing is a mechanism to determine students' abilities in the language earlier on in their university studies, at a time when they would still be able to improve. In other words, there has been a perceived need for a formative assessment of the English proficiency of students while at university that could both measure students' development in the language and offer diagnostic information to help them to improve. The Diagnostic English Language Tracking Assessment (DELTA) was developed to address this need.

DELTA is considered to be an assessment of English language proficiency. According to Bachman and Palmer's (1996) conceptualisation, a frame of reference such as a course syllabus, a needs analysis or a theory of language ability is required to define the construct to be assessed. As DELTA is not tied to any particular syllabus and is designed for use in different institutions, it is therefore based on a theory of language ability, namely that presented in Bachman and Palmer (1996), which was derived from that proposed by Bachman (1990). Under their theory or framework, language ability consists of language knowledge and strategic competence. As the format of DELTA is selected response, it is considered that strategic competence does not play a major part in the construct and so language knowledge (including organisational knowledge – grammatical and textual, and pragmatic knowledge – functional and sociolinguistic) (Bachman and Palmer 1996) is regarded as the underlying construct. Given the academic English *flavour* of DELTA, it should be regarded as an academic language proficiency assessment, as students have to apply their knowledge of academic English to be able to process the written and spoken texts and answer the test items.

DELTA consists of individual multiple-choice tests of listening, vocabulary, reading and grammar (with a writing component under development). The reading, listening, and grammar items are text-based, while the vocabulary items are discrete. Despite the fact that a multiple-choice item format limits the items to assessing receptive aspects of proficiency, this format was chosen to allow for immediate computerised marking. The Assessment lasts 70 min. Each component (except vocabulary) consists of a number of parts, as shown in Table 5.2.

The listening and reading components consist of four and three parts respectively and the grammar component two parts. The DELTA system calculates the total number of items in these three components and then adds items to the vocabulary component such that the total number of items on the assessment equals 100.

DELTA uses discrete items which test identified subskills of listening, reading, grammar and vocabulary. The subskills are written in accessible language in the DELTA student report. In addition, Rasch measurement is used to calibrate items and generate the reports. Student responses to the test items are exported from the DELTA system and imported into WINSTEPS v3.81 (Linacre 2014), a Windows-based Rasch measurement software program. After analysis, students' DELTA

such that final-year undergraduates were funded to take IELTS and the results used by the Government as a measure of the effectiveness of English language enhancement programmes at the Government-funded institutions. In addition, one institution, the Hong Kong Polytechnic University, requires its graduating students to take its own Graduating Students' Language Proficiency Assessment (GSLPA).

Table 5.2 The structure of DELTA

Component	Parts	Composition	Difficulty	Time allowed
Listening	Part 1	1 Recording + 4–6 items	Easier	20–25 min
	Part 2	1 Recording + 6–8 items	↓	
	Part 3	1 Recording + 6–8 items		
	Part 4	1 Recording + 6–8 items	More difficult	
Vocabulary		20–25 Items	A range	45–50 min
Reading	Part 1	1 Text + 4–6 items	Easier	
	Part 2	1 Text + 6–8 items	↓	
	Part 3	1 Text + 6–8 items	More difficult	
Grammar	Part 1	1 Text + 10–15 items	A range	
	Part 2	1 Text + 10–15 items		

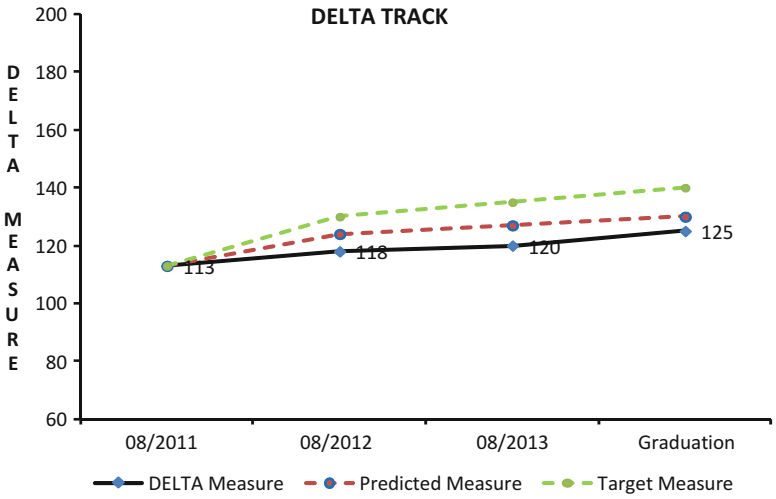


Fig. 5.1 DELTA track

Measures and item calibrations are returned to the DELTA system and the reports are generated. DELTA Measures are points on the DELTA proficiency scale, set at 0–200, which allows progress (or growth) to be tracked each time a student takes the DELTA (see Fig. 5.1).[6]

With the aid of the report that students receive after taking DELTA[7] plus the kind of programmes and activities described, DELTA has been employed at four universities in Hong Kong to help raise English proficiency and motivate students to continue to maintain or (hopefully) raise their proficiency throughout their time at

[6] For a detailed description of the development and implementation of DELTA, see Urmston et al. (2013).

[7] For a sample DELTA report, see Appendix.

Table 5.3 Numbers of students taking DELTA since 2011

	No. of students taking DELTA					
	2011	2012		2013		
University	First attempt	First attempt	Second attempt	First attempt	Second attempt	Third attempt
Hong Kong Polytechnic University	398	728	0	889	24	0
City University of Hong Kong	1645	2152	534	827	7	2
Lingnan University	488	1156	0	586	444	0
Baptist University	0	0	0	77	0	0
Total	2531	4036	534	2379	475	2

university. DELTA was developed as a collaboration between the Hong Kong Polytechnic University, City University of Hong Kong and Lingnan University, and was first launched on an operational scale in the 2011–2012 academic year. A fourth institution, Baptist University, began using DELTA on a small scale in 2013. The numbers of students taking DELTA at each university are shown in Table 5.3.

There were various reasons why the numbers of students taking DELTA at first or subsequent attempts differed from university to university, including changes in policy, which have either made DELTA a compulsory requirement for students or reversed a decision to do so. A fuller description of the reasons behind such decisions is beyond the scope of this chapter and the reader is directed to Urmston et al. (2013).

For a diagnostic test to be effective, it should emphasise the identification of learners' weaknesses (Alderson 2005; Buck 2001). That is, the test should be as comprehensive as possible and able to provide detailed information about the subskills involved in accomplishing tasks or answering test items. This is what DELTA does, while adhering to Alderson et al.'s (2015, p. 21) (tentative) principles for diagnostic second or foreign language (SFL) assessment, i.e.

> The first principle … is that it is not the test that diagnoses, it is the user of the test.

DELTA accepts this principle and adheres to it by providing an interactive report that users – mainly students and teachers – can then interpret and use in their own way. The DELTA report provides the evidence and the test users then must decide on the action to take to gain the most from the report, within the constraints of time and motivation that they have.

> The second principle is that instruments themselves should be designed to be user-friendly, targeted, discrete and efficient in order to assist the teacher in making a diagnosis.

The development of DELTA placed user-friendliness as one of its guiding principles. Though DELTA is an online assessment, it cannot be assumed that all test takers are equally computer-literate. Difficulty in using the assessment interface

will introduce construct-irrelevant variance and should be avoided. DELTA is also targeted to its test takers, initially at a generic level and subsequently more closely, as test-takers' previous performance is taken into account when items are selected from the item bank to make up the version of DELTA that the test taker is given. DELTA is made up of discrete items[8] and all procedures are carried out efficiently.

> The third principle is that the diagnostic assessment process should include diverse stakeholder views, including learners' self-assessments.

DELTA is used in different contexts and to varying degrees different stakeholders' views are taken into account. At Lingnan University, for example, students must complete an independent learning plan to supplement their classes in the language. This involves a degree of self-assessment. At the Hong Kong Polytechnic University, students enrolling on the Excel@English Scheme (EES) work with teacher mentors to further the diagnostic process which begins with DELTA.

> The fourth principle is that diagnostic assessment should ideally be embedded within a system that allows for all four diagnostic stages: (1) listening/observation, (2) initial assessment, (3) use of tools, tests, expert help, and (4) decision-making …

Again, programmes such as the EES, in which DELTA is embedded, enable the diagnosis to continue as a process beyond the taking of the assessment itself. At Baptist University, students identified as in need of language support take DELTA as a component of the English Language Mentoring Service (ELMS). According to the website of the Language Centre at Baptist University:

> ELMS … aims to provide students with a supportive environment conducive to English language learning in their second semester at university. The content will be discussed and negotiated between the students and lecturers, taking into consideration the wishes and needs of the students and the expertise and advice of the lecturer. Students will also be encouraged to learn relevant topics at their own pace and in their own time, and according to their individual needs and priorities (http://lc.hkbu.edu.hk/course_noncredit.php).

The importance of including DELTA in a learner-teacher negotiated programme of diagnosis and intervention cannot be understated and this is encapsulated in the fifth of Alderson et al's (2015) principles.

> The fifth principle is that diagnostic assessment should relate, if at all possible, to some future treatment.

Wherever feasible this is the case with DELTA, though students are able to take the assessment on a voluntary basis and whether or not there is future treatment depends on the student. In such cases, students are strongly encouraged to make use of the learning resources provided through links embedded into the DELTA report and to seek advice from an instructor or language advisor.

[8] In fact for listening, reading and grammar, as DELTA adopts a testlet model, its items cannot be considered as totally discrete, though for purposes of item calibration and ability measurement, unidimensionality is assumed.

2 The Study

This chapter reports on a study of the 475 students who took DELTA for the second time in 2013. These were students for whom it would be possible to determine whether their English proficiency had changed over one academic year at university in Hong Kong. The study aimed to answer the following research questions:

1. Can DELTA reliably measure difference in students' English language proficiency over a 1-year period?
2. Was there a difference in the students' proficiency, i.e. their DELTA Measures, between their first and second attempt of the DELTA?
3. What might account for this difference (if any) in DELTA Measures?
4. Does the DELTA have any impact on students' language proficiency development and if so, how?

3 Methodology

The study attempts to answer the first two research questions by determining empirical differences in DELTA Measures of the 475 students over the two attempts at DELTA using statistical methods described in the following sections. To answer the third and fourth questions, a questionnaire was administered to gain information on students' past and current language learning activities, experiences and motivations, their perceptions of the ability of DELTA to measure their English proficiency, and the usefulness of the DELTA report. The questionnaire was delivered online through Survey Monkey and students were asked to complete it after they had received their second attempt DELTA report. A total of 235 students responded to the questionnaire. Of these 235, in-depth focus group interviews were conducted with eight students to determine the extent to which DELTA has any impact on students' language learning while at university. Content analysis of the data (Miles and Huberman, 1994; Patton 2002) was undertaken using QSR NVivo 10.

4 Data Analysis and Results

4.1 Examining the Psychometric Quality of the Test Items Across Time

Overall, 2244 test items (1052 in 2012 and 1192 items in 2013) were used to assess the reading and listening ability as well as grammar and vocabulary knowledge of the students who took DELTA in 2012 and 2013. One thousand and five items were common and used to link the test takers. The psychometric quality of the items

Table 5.4 Reliability and separation indices of the DELTA components across time

	2012		2013	
Test	Reliability	Separation	Reliability	Separation
Reading	.97	5.50	.96	4.70
Listening	.99	8.37	.98	6.65
Grammar	.97	5.47	.92	3.48
Vocabulary	.99	13.28	.99	8.27

across time was examined using Rasch measurement. Initially, the items administered in 2012 were linked and fitted to the Rasch model. The reliability and separation statistics, item difficulty, and infit and outfit mean square (MNSQ) coefficients were estimated. Item difficulty invariance as well as reliability across time were also checked. Item reliability in Rasch measurement is an index for the reproducibility of item difficulty measures if the items are administered to a similar sample drawn from the same population. Separation is another expression of reliability that estimates the number of item difficulty strata.

Infit and outfit MNSQ values are chi-square statistics used for quality control analysis and range from zero to infinity; the expected value is unity (1), but a slight deviation from unity, i.e., between 0.6 and 1.4, still indicates productive measurement in the sense that the data is likely not affected by construct-irrelevant variance.

To estimate item difficulty measures, each test component was analyzed separately using WINSTEPS v3.81 (Linacre 2014). In each analysis, the item difficulty measures were generated by first deleting misfitting items and the lowest 20 % of persons (students) to ensure the best calibration of items for each component. These calibrations were then used to generate person measures.

Table 5.4 presents the reliability and separation indices of the reading, listening, grammar and vocabulary components across time. The components maintain their discrimination power; for example, the reliability and separation coefficients of the reading component in 2012 were .97 and 5.50, respectively, and the reliability index in 2013 was highly similar (.96); the separation index, if rounded, indicates the presence of approximately five strata of difficulty across the two time points. We also note similarities between separation and reliability estimates of the other DELTA components across the 2 years. A seemingly large difference exists between the separation statistics of the vocabulary test across the 2 years, despite their equal reliability estimates. The discrepancy stems from the nature of reliability and separation indices: whereas the near-maximum reliability estimate (.99) is achieved, the separation index has no upper bound limit and can be any value equal to or greater than eight, depending on the sample size and measurement error (Englehard 2012). Overall, there is evidence that the reliability of the components did not drop across time. In addition, the infit and outfit MNSQ values of the test items all fell between 0.6 and 1.40, indicating that the items met the requirements of the Rasch model and it was highly unlikely that construct-irrelevant variance confounded the test data.

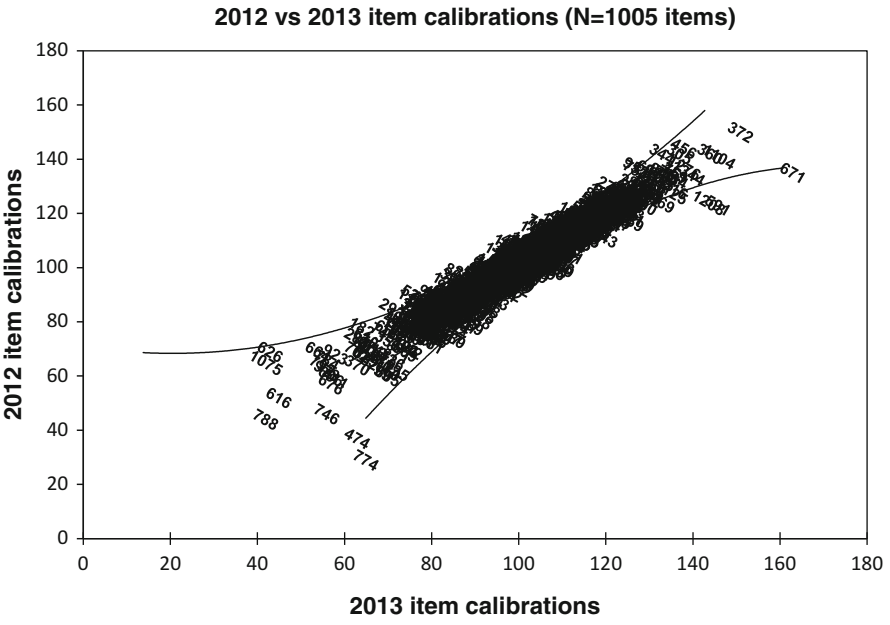


Fig. 5.2 Item calibrations plotted across time

4.2 Invariance Over Time

We inspected the invariance of item difficulty over time by plotting the difficulty of items in 2012 against that in 2013. Figure 5.2 presents the related scatterplot of the 1005 items which were common across administrations. As the figure shows, almost all items fall within the 95 % two-sided confidence bands represented by the two lines, suggesting that there was no significant change in item difficulty over time. This can be taken as evidence for invariant measurement since the psychometric features of the test items are sustained across different test administrations (Engelhard 2012).

4.3 Examining the Development of Test Takers Over Time

After ensuring that the test items had stable psychometric qualities, we performed several bootstrapped paired sample t-tests with 1000 bootstrap samples to determine whether there was a significant difference between the mean DELTA measures of the students across time. Bootstrapping was used to control and examine the stability of the DELTA results, approximate the population parameters, and estimate the true confidence intervals (DiCiccio and Efron 1996). This test initially estimates the correlation between the DELTA measures over time. Table 5.5 presents the correlation coefficients of the DELTA measures in 2012 and 2013. Except in the case of the overall measures, the correlation coefficients are below .700, meaning that,

Table 5.5 Bootstrapped correlation coefficients across time

	Correlation	*SD*	95 % confidence interval	
			Lower	Upper
Listening 2012 and listening 2013	.449	.036	.378	.516
Vocabulary 2012 and vocabulary 2013	.572	.030	.505	.630
Reading 2012 and reading 2013	.394	.037	.319	.470
Grammar 2012 and grammar 2013	.429	.037	.356	.497
Overall measure 2012 and overall 2013	.703	.027	.645	.754

Bootstrap results are based on 1000 bootstrap samples

Table 5.6 Bootstrapped paired sample t-test results

	Mean	SD	*p* value (2-tailed)	95 % confidence interval	
				Lower	Upper
Listening2012 – listening2013	−.22425	.03936	.001	−.30364	−.14318
Vocabulary2012 – vocabulary2013	−.28029	.04302	.001	−.36058	−.19415
Reading2012 – reading2013	.00509	.03779	.888	−.06465	.08404
Grammar2012 – grammar2013	.00728	.03979	.855	−.07086	.08808
PersonMeasure2012 – person measure2013	−.1774608	.0179402	.001	−.2115781	−.1403957

Bootstrap results are based on 1000 bootstrap samples

on average, the rank-ordering of students across times one and two tended to be rather dissimilar (Field 2005) (e.g. low/high measures in 2012 are not highly associated with low/high measures in 2013). This suggests that there might have been an increasing or decreasing trend for the majority of the students who took the DELTA. The bootstrapped correlation also provides 95 % confidence intervals, indicating that the true correlations (the correlation of the components in the population) fall between the lower and upper bands. For example, the estimated correlation between the listening measures of the students in 2012 and 2013 was .449, with lower and upper bands of .378 and .516 respectively. This suggests that the estimated correlation is close to the mean of the bands and is therefore highly reliable.

The bootstrapped paired sample t-test results are presented in Table 5.6. While the listening, vocabulary and overall test measures significantly increased across time, as indicated by the significant mean differences ($p < 0.001$), the reading and grammar measures had no significant increase.

To relate this evidence of increase in DELTA measures to growth in terms of English language proficiency of the students, we took as a parameter a 0.5 logit difference in the DELTA scale.[9] This was then used as the cut-off point to determine

[9] A difference of 0.5 logits on a scale of educational achievement is considered statistically significant in educational settings based on OECD (2014) results of an average of 0.3–0.5 annualised score point change reported by the PISA 2012 test. The PISA scale is from 0 to 500.

Table 5.7 Profile of second attempt DELTA candidates 2013

	Overall measure	Listening	Vocabulary	Reading	Grammar
Growers (+0.5 logit)	103	171	194	120	124
Sustainers (minimal or no change)	352	217	192	228	225
Decliners (−0.5 logit)	20	87	89	127	126

Table 5.8 Students' perception of improvement in component skills

	Listening	Reading	Grammar	Vocabulary
Do you think you have made improvement in the different language skills during the year between taking DELTA?	60 %	69 %	55 %	55 %

growers (>+0.5 logit difference), *sustainers* (measures have minimal change or none at all), and *decliners* (>−0.5 logit difference). Table 5.7 summarises the total number of growers, sustainers, and decliners from 2012 to 2013.

The findings show that there were noticeably more growers than decliners in listening and vocabulary and in the overall measure, while there was no significant difference in reading and grammar.

4.4 Students' Perception of Improvement While at University

The survey results of students' perception of improvement, however, are a little inconsistent with the quantitative results. Of the four components, reading was the skill that most of the students felt they had made improvement in after their first year of studies, followed by listening, grammar and vocabulary (Table 5.8).

This result is consistent with the findings of Evans and Morrison (2011), who found that first year university students in Hong Kong face "a relentless diet of disciplinary reading and listening" (p. 203), and so it is perhaps unsurprising that they feel that these skills have improved most, due to practice more than anything else, as there is little in the way of direct instruction in these skills.

Digging more deeply to look at whether there were particular subskills in which students felt they had improved the most, although there was the feeling that they had improved in all subskills, there was limited evidence that they considered they had improved in any one more than another. Figures 5.3 and 5.4 present the students' perceptions of improvement in reading and listening and seem to show to a small extent that students felt there was some improvement in understanding main ideas and supporting details, at least in listening.

In grammar, students felt that they had made more improvement in 'grammatical accuracy in writing' than in speaking (Fig. 5.5).

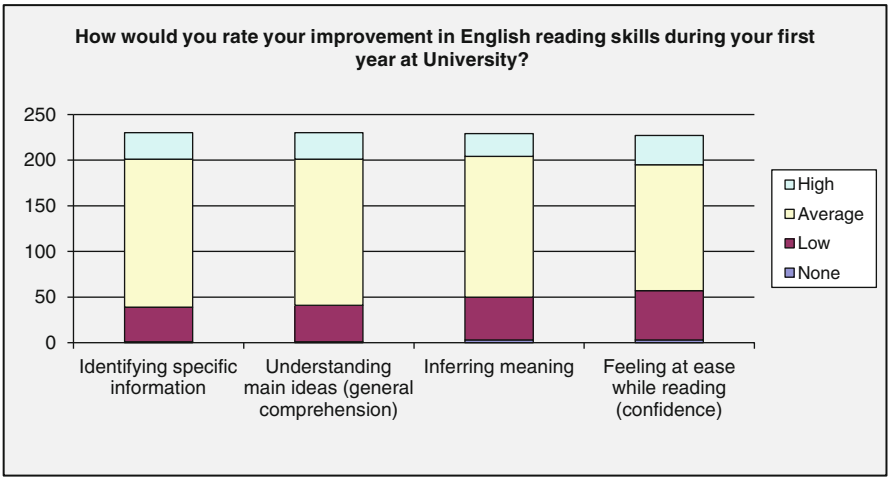


Fig. 5.3 Students' perception of improvement in reading skills

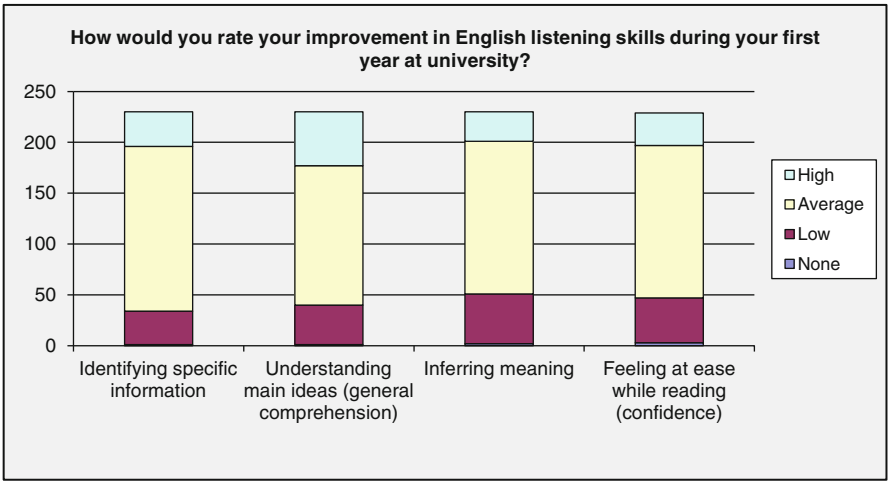


Fig. 5.4 Students' perception of improvement in listening skills

The disproportionate emphasis on writing compared to speaking makes this result unsurprising too. Teachers in Hong Kong tend to give extensive feedback on writing, focusing particularly on grammatical accuracy and error correction (Lee and Coniam 2013). In terms of vocabulary, slightly more improvement was reported in understanding unknown words from context, though again the students felt that they had improved in all areas (Fig. 5.6).

These survey results are also not surprising considering the students' responses to questions on their use of English inside and outside campus. According to the survey, a large majority of the students (around 90 %) used spoken English from

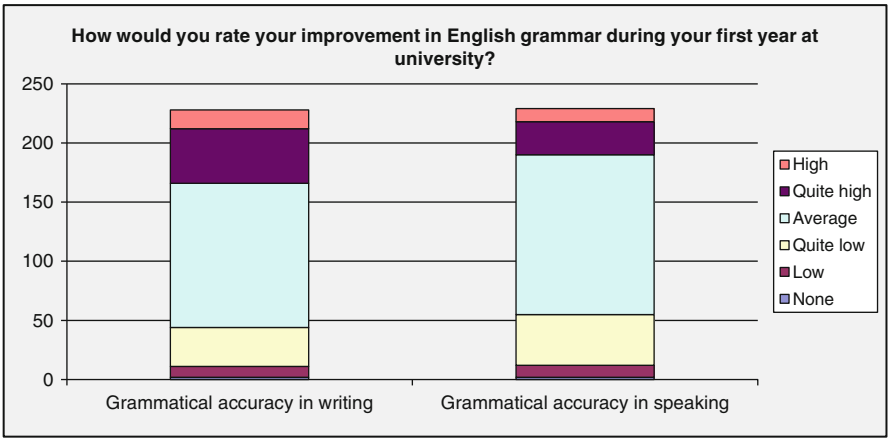


Fig. 5.5 Students' perception of improvement in grammar

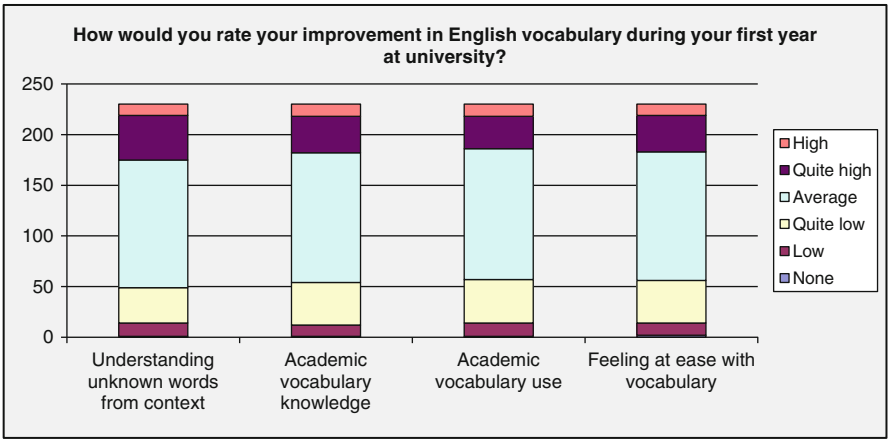


Fig. 5.6 Students' perception of improvement in vocabulary

'often' to 'always' in compulsory academic settings such as lectures, seminars and tutorials, whereas English was rarely used with peers in their residence hall activities and other extracurricular activities. In addition about 50% said they used English in their part-time jobs (often involving tutoring or sales work) (Table 5.9).

These results suggest that the students felt that they had made improvement in their ability to listen for main ideas and supporting details and understand unknown words from context because of the need to use English in academic settings where they are required to read extensively, and listen to lectures in English for around 20 h a week.

Table 5.9 Students' self-reported frequency of use of spoken English (%)

How often do you use English in the following situations?	Always	Almost always	Often	Some-times	Rarely	Never
In lectures	27	**32**	25	11	4	1
In seminars/tutorials	24	**36**	29	9	1	1
Outside class with other students	6	8	**30**	23	26	2
Residence hall activities	4	8	20	20	**40**	10
Extra-curricular activities	4	9	22	25	**31**	9
Part-time job	6	11	**27**	25	25	6

Table 5.10 Students' self-reported use of written English (%)

How often do you do the following activities in English?	More than 5 hrs a day	About 3–5 hrs a day	About 1–3 hrs a day	Less than an hour a day	Never
Online messaging (including WhatsApp, Skype, MSN etc.)	8	13	35	**39**	4
Email	2	10	25	**59**	4
Social media (including Facebook or Twitter etc.)	5	13	36	**39**	7
Internet chat forums, blogs or homepages	1	8	22	**50**	19
Browsing websites	3	9	27	**54**	7
Reading books/magazines/ newspapers	2	7	30	**50**	10
Listening to music	8	12	36	**41**	3
Watching TV/movies	5	12	31	**49**	3

Of other activities, English was used by more than half the students for email, online messaging, social media and listening to music for more than an hour a day (Table 5.10).

The perceptions of improvement or lack of it are further explained by the students' motivations to learn English while at university. Survey results revealed that students' main reasons for English language learning were factors such as meeting a course requirement, eligibility to participate in exchange programmes, importance of English for future career, and encouragement from teachers and parents. However, lack of confidence in their ability to learn English and feelings of anxiety while learning a language continued to be hindrances. These results suggest that English language learning while at university is mainly for pragmatic reasons, i.e. the need to use English for academic purposes.

4.5 Students' Perception of the Impact of DELTA on their English Language Learning Habits

In order to determine how students perceive the impact of DELTA on their English language learning habits, eight of the students from Lingnan University were asked to participate in focus group interviews to elaborate on their perceptions of DELTA and its impact on their English learning. At Lingnan University, students use their DELTA report as input for the independent learning component of their compulsory English language enhancement course, English for Communication II. The independent learning component accounts for 20 % of the course grade and many students do use their DELTA reports for diagnosis, i.e. to help identify areas of relative weakness, formulate learning plans or pathways and work on these in their independent learning.

Three *growers* and five *sustainers* participated in the focus group interviews. All of the students claimed that DELTA was able to reflect their English proficiency in that the DELTA report accurately reported their strengths and weaknesses. All of them used the report as a first step to improve their English proficiency. What distinguished the growers from the sustainers, however, was how they approached their own language learning. First, a quote from one of the growers:

> *I tried* [using the independent learning links in the DELTA report] *when I was in year one, semester one but I stopped trying it because I have my own way of learning English, which is like in last year, my semester one, I listened to TED speech. I spent summer time reading ten English books and made handful notes. I also watched TVB news* [a local English-language TV station] *online to practice my speaking. I also watched a TV programme. I used to use the advanced grammar book and there is a writing task, I forget the name, I bought it. It helped to improve my English. It's really a good book, it helped me to improve my grammar and writing skills. So people have different ways to learn English. I've found my way to learn English. I think these websites may be useful to someone, but not to me.* Tony (grower)

Tony described how he independently developed his own language learning without help from others. He made an effort to surround himself with English (e.g., listening to talks, reading books, doing grammar activities) because he believed it was the only way to improve. He did not think the independent learning links in the DELTA report were particularly useful to him as they did not suit his style of learning.

On the other hand, two of the sustainers took a different approach.

> *I felt curious to use this website to improve my listening. This year is my second time, I don't consider it important.* Sia (sustainer)
>
> *I think my instructor wouldn't know the details of the report. He just said, "you refer to the DELTA report to decide which skill you want to improve when planning your independent learning". Now when I see this report shows my vocab is weakest, which I agree, I feel the DELTA, in addition to helping you to make an independent learning plan, shows us what skill I have problem with. I think what is more important is that the instructor can tell in you detail which skill is weak and has to be worked on.* Elsa (sustainer)

Table 5.11 Top English activities by growers and sustainers that helped improve their English

1	Reading in English (fiction, non-fiction, magazines)
2	Using self-access centre (Speaking and/or Writing Assistance Programme)[a]
3	Listening to lectures[a]
4	Watching TV shows or programmes
5	Text messaging
6	Talking to exchange students (inside or outside the classroom)
7	Academic reading (journal articles, textbooks)[a]
8	Using dictionary to look for unknown words
9	Listening to and/or watching TED talks
10	Doing grammar exercises
11	Listening to music
12	Test preparation
13	Watching YouTube clips
14	Watching movies
15	Doing online activities
16	Attending formal LCE classes[a]
17	Memorising vocabulary
18	Joining clubs and societies
19	Reading and writing emails
20	Exposure to English environment

[a]Study-related activities

Sustainers were similar to growers in that they did not find the independent learning links provided in the DELTA report useful for their learning. However, they required further guidance from teachers to improve their English. They felt that the DELTA report was useful and accurately reflected their strengths and weaknesses but they attributed their lack of development to not having support from teachers to show them what the next step in their language learning should be. This confirms the survey finding that lack of confidence in their ability to learn English was a hindrance to further development, as well as supporting Alderson et al.'s (2014) second principle for diagnostic assessment, that teacher involvement is key.

The participants were also asked to describe the top activities that they thought helped them improve various aspects of their English. Table 5.11 lists the top 20 activities that growers and sustainers specifically thought were useful in their English language growth.

Surprisingly, only three of the top ten activities are study-related (listening to lectures, using self-access centre and academic reading) and the rest are all non-study-related activities. Reading in English was the most popular activity followed by the use of services offered by the self-access centre and finally listening to lectures and watching TV shows or programmes in English. These results suggest that

if students want to improve their English, they clearly have to find activities that suit their learning styles, and this in turn will motivate them to learn. As Tony said,

> *So I think it's very important when you think about your proficiency - if you're a highly motivated person then you will really work hard and find resources to improve your English. But if you're like my roommate, you don't really work hard in improving English, then his English proficiency skills will be just like a secondary school student. Seriously.* Tony (grower)

Clearly then, as concluded by Alderson (2005), it is the (informed) intervention in the learning process that is the most essential contribution that diagnostic testing can make. The developers of DELTA have worked hard to provide the support that students need during their development, including links to learning websites and online resources, teacher mentoring programmes and extracurricular activities, to help motivate students to continue to engage in the language learning process.

5 Discussion

The first of our research questions asked whether the diagnostic testing instrument used, DELTA, can reliably measure difference in students' English language proficiency over a 1-year period. Overall, the results of the psychometric analysis provided fairly strong support for the quality of the four component tests (listening, reading, grammar and vocabulary). In addition, the bootstrapped paired sample t-test results indicated that there was a statistically significant difference between students' performance across time. In other words, DELTA can be used to measure differences in English language proficiency over a 1-year period.

Secondly, there was a difference shown in some students' proficiency, i.e. their DELTA Measures, between their first and second attempts of the DELTA. Inevitably, some students improved or grew while others actually showed regression or decline. In most cases, though, there was no difference measured. Results seemed to indicate an overall increase in proficiency of the group in terms of numbers of growers being greater than the numbers of decliners, which would no doubt please university administrators and programme planners. More specifically, there were more growers than decliners in listening and vocabulary, while reading and grammar saw no discernible change. Such information again is useful for programme planners and teachers in that they can look at which aspects of their English language provision they need to pay more attention to.

In seeking what might account for this difference in DELTA Measures, we have looked at students' reported English-use activities. Time spent in lectures, seminars and tutorials requiring them to listen in English seems to have impacted their proficiency in this skill, while their self-reported attention to academic reading seems to have improved their academic vocabulary to a greater extent than their reading

skills. Indications are that students who do show growth are those that adopt their own strategies for improvement to supplement the use of the language they make in their studies.

Qualitative results suggest that DELTA has impact as students report that it is valuable as a tool to inform them of their strengths and weaknesses in their English proficiency. For those required to create independent learning plans, DELTA reports are the first source of information students rely on. The real value of DELTA, however, is the tracking function it provides. Interviews with growers and sustainers suggest that those students who want to improve their proficiency obviously do more than the average student; these students are fully aware of their learning styles and seek their own learning resources and maximize these. Thus, DELTA's tracking function serves to validate the perception that their efforts have not been in vain. This suggests that perhaps DELTA should be part of a more organised programme which helps students identify learning resources that suit their learning styles and needs and involves the intervention of advisors or mentors. An example of this is the Excel@English Scheme (EES) at Hong Kong Polytechnic University mentioned previously. This scheme integrates DELTA with existing language learning activities as well as custom-made learning resources and teacher mentoring. It allows for student autonomy by providing the support that is clearly needed.

6 Conclusion

This chapter has described how a diagnostic assessment can be used to inform and encourage ESL students' development in English language proficiency as support for them as they progress through English-medium university studies. The assessment in question, the Diagnostic English Language Tracking Assessment (DELTA) has been shown to provide reliable measures of student growth in proficiency, while the diagnostic reports have proved to be a useful starting point for students in their pursuit of language development. What has become clear, though, is that the diagnostic report alone, even with its integrated language learning links, is not enough and students need the support of teachers to help them understand the results of the diagnostic assessment and provide the link to the resources they can use and the materials that are most appropriate for them, given their needs and learning styles. Clearly a bigger picture needs to be drawn to learn more about how a diagnostic assessment like DELTA can impact language development, and this will be possible as more students take the assessment for a second, third or even fourth time. Language proficiency development is a process and it is to be hoped that for university students, it is one that is sustained throughout their time at university.

7 Appendix

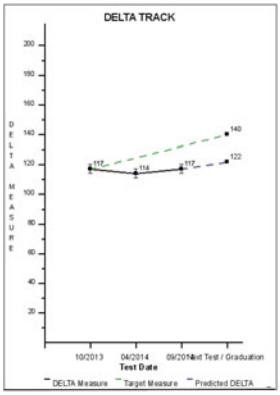
THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

DELTA Diagnostic English Language Tracking Assessment

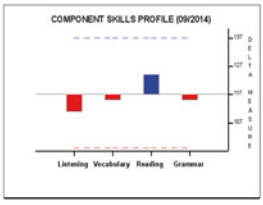
Diagnostic English Language Tracking Assessment
Candidate Report

Name: **Student No:** **Date:**

This is a report of your performance in the Diagnostic English Language Tracking Assessment (DELTA). Each time you take the DELTA, you will receive a DELTA Measure and a diagnostic report. The reports are cumulative, so that you can track your progress in improving your English.



The DELTA Track shows your English proficiency calculated from your performance on the DELTA tests. Each time you take the tests, your DELTA Measure is plotted to show your progress.



The Component Skills Profile above shows the contribution that the scores that you attained on each component has made to your DELTA Measure. Components below the line indicate areas of relative weakness. Click on each component for more details.

Last updated:

Component Diagnostic Report

The four reports below show your performance on each of the four components in the DELTA. They show a description of the subskills tested by each of the items you did, in descending order of difficulty. Your proficiency level as indicated by your DELTA Measure is also shown. Items below the line of your proficiency level are those that you would be expected to answer correctly. The items that are highlighted indicate the subskills that you should focus on in your English language learning. Click on each subskill to learn more about it.

Listening

Difficulty		Subskills tested	Text Type	Theme
More ↑	×	Understanding information and making an inference	TV/Radio interviews	Leisure and entertainment
	√	Interpreting a word or phrase as used by the speaker		
	×	Identifying specific information	Presentations and lectures	Business and marketing
	√	Understanding information and making an inference		
				DELTA 117
	×	Identifying specific information	Presentations and lectures	Business and marketing
	√	Understanding information and making an inference		
	√	Understanding information and making an inference		
	×	Identifying specific information	News reports	Environmental issues
	×	Identifying specific information	News reports	Environmental issues
	√	Understanding information and making an inference		
	×	Identifying specific information	TV/Radio interviews	Leisure and entertainment
	×	Identifying specific information	Presentations and lectures	Business and marketing
	√	Identifying specific information		
	√	Identifying specific information		
	√	Understanding main ideas and supporting ideas		
	√	Identifying specific information		
	√	Identifying specific information		
	√	Understanding information and making an inference		
	×	Identifying specific information	News reports	Environmental issues
	√	Understanding main ideas and supporting ideas		
Less ↓	√	Identifying specific information		

The report for Listening above indicates subskills to work on. The online version of the report provides links to specific learning resources. You can find resources for improving your listening skills in general at http://elc.polyu.edu.hk/cill/listening/.

Last updated:

Vocabulary

		Academic Word Sublist	Words for revision
More ↑	×	AWL Sublist 4	unparalleled
	×	AWL Sublist 8	conformance
	×	AWL Sublist 9	erosion
	√	AWL Sublist 3	
	×	AWL Sublist 7	inferred
	×	AWL Sublist 5	discrete
			DELTA 117
	√	AWL Sublist 9	
	×	AWL Sublist 1	derived
	×	AWL Sublist 3	unconventional
	√	AWL Sublist 10	
	×	AWL Sublist 9	devoted
	√	AWL Sublist 6	
Difficulty	×	AWL Sublist 6	aggregated
	√	AWL Sublist 9	
	√	AWL Sublist 10	
	√	AWL Sublist 9	
	√	AWL Sublist 8	
	×	AWL Sublist 4	adequate
	√	AWL Sublist 4	
	√	AWL Sublist 9	
	√	AWL Sublist 7	
	√	AWL Sublist 1	
	√	AWL Sublist 9	
	√	AWL Sublist 9	
	√	AWL Sublist 1	
Less ↓	√	AWL Sublist 3	
	√	AWL Sublist 10	

In the report for Vocabulary above, the items are drawn from the Academic Word List (AWL). The AWL is divided into 10 sublists of words used in academic English. Sublist 1 consists of the most frequent words. Sublist 2 contains the next most frequent and so on. The online version of the report provides links to specific learning resources. You can find more information about the AWL at http://elc.polyu.edu.hk/cill/vocabulary/.

Last updated:

Reading

		Subskills tested	Text type	Theme
More ↑	×	Identifying specific information	Feature articles	Technology
	×	Inferring the writerâ€™s reasoning	Natural sciences	Feature articles
	×	Understanding information and making an inference	Feature articles	Technology
	√	Understanding main ideas and supporting ideas		
	√	Interpreting a word or phrase as used by the writer		
	√	Interpreting an attitude or intention of the writer		
				DELTA 117
	√	Understanding information and making an inference		
	√	Understanding main ideas and supporting ideas		
	×	Identifying specific information	Feature articles	Technology
	√	Identifying specific information		
	√	Understanding main ideas and supporting ideas		
Difficulty	×	Understanding information and making an inference	Fiction	Daily life
	×	Understanding main ideas and supporting ideas	Fiction	Daily life
	√	Understanding main ideas and supporting ideas		
	√	Identifying specific information		
	√	Identifying specific information		
	√	Understanding main ideas and supporting ideas		
	√	Interpreting an attitude or intention of the writer		
	√	Identifying specific information		
	√	Understanding information and making an inference		
	√	Understanding main ideas and supporting ideas		
Less ↓	√	Interpreting an attitude or intention of the writer		

The report for Reading above indicates subskills to work on. The online version of the report provides links to specific learning resources. You can find resources for improving your reading skills in general at http://elc.polyu.edu.hk/cill/reading/.

Last updated:

Grammar

		Subskills tested	Example
More ↑	×	Determiner	She had a long holiday in (the Asia) ASIA last year.
	×	Determiner	She had a long holiday in (the Asia) ASIA last year.
	√	Preposition	
	×	Preposition	It was late so he came home (on) IN a taxi.
	×	Preposition	It was late so he came home (on) IN a taxi.
	×	Past perfect tense	By the time I got up, he (ate) HAD EATEN all the sausages.
	×	Participle	They say she died of a (broke) BROKEN neck.
			DELTA 117
	×	Simple present tense	It always (rained) RAINS at this time of year.
	√	Prepositional phrase	
	×	Word form	It is a very (comforting) COMFORTABLE chair.
	√	Phrasal verb	
	√	Subject and verb agreement	
	√	Determiner	
Difficulty	√	Participle	
	√	Preposition	
	×	Transitive verb	She (surprised) SURPRISED THE AUDIENCE with the quality of her presentation.
	√	Verb formation	
	×	Voice	The law should (amend) BE AMENDED immediately.
	√	Appositive clause	
	√	Voice	
	√	Connective	
	×	Relative clause	The main idea for the project, (who) WHICH is led by Professor Chan, is to design a better system.
	√	Demonstrative pronoun	
	√	Relative clause	
	√	Indefinite pronoun	
	√	Adverb	
	√	Prepositional phrase	
	√	Voice	
Less	√	Determiner	
	√	Comparative	

The report for Grammar above indicates subskills to work on. In the second column, examples are provided for the incorrect items. The error is in brackets and the correct form is in capital letters. The online version of the report provides links to specific learning resources. You can find resources for improving your grammar skills in general at http://elc.polyu.edu.hk/cill/grammar/.

Last updated:

Overall Performance

Your Component Skills Profile suggests that you should prioritise your English language learning as follows:

1. Listening
2. Grammar
3. Vocabulary
4. Reading

You should aim to improve your proficiency by focusing in particular on those areas in which you have shown weakness, making use of the links provided to relevant learning resources. You may also seek the advice and/or guidance of an EES Mentor or English Language Centre teacher to make the best use of this report.

The next time you take the DELTA, it will be targeted to your proficiency level, enabling you to demonstrate the progress that you have made.

Thank you for taking the DELTA and good luck in your English language learning.

Language Testing Unit
English Language Centre

ENGLISH LANGUAGE CENTRE

Last updated:

References

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Alderson, J. C., Brunfaut, T., & Harding, L. (2015). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. *Applied Linguistics, 36*(2), 236–260.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7–71.

Buck, G. (2001). *Assessing listening*. New York: Cambridge University Press.

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science, 11*, 189–228.

Engelhard, G. (2012). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York: Routledge.

Evans, S., & Green, C. (2007). Why EAP is necessary: A survey of Hong Kong tertiary students. *Journal of English for Academic Purposes, 6*(1), 3–17.

Evans, S., & Morrison, B. (2011). Meeting the challenges of English-medium higher education: The first-year experience in Hong Kong. *English for Specific Purposes, 30*(3), 198–208.

Field, A. (2005). *Discovering statistics using SPSS* (2nd ed.). London: Sage.

Lee, I., & Coniam, D. (2013). Introducing assessment for learning for EFL writing in an assessment of learning examination-driven system in Hong Kong. *Journal of Second Language Writing, 22*(1), 34–50.

Linacre, J. M. (2014). *Winsteps: Rasch model computer program (version 3.81)*. Chicago: www.winsteps.com

Miles, M. B., & Huberman, M. A. (1994). *Qualitative data analysis: An expanded sourcebook*. Thousand Oaks: Sage.

Miller, L., & Gardner, D. (2014). *Managing self-access language learning*. Hong Kong: City University Press.

OECD. (2014). *Pisa 2012 results in focus: what 15-year-olds know and what they can do with what they know.* The Organisation for Economic Co-operation and Development (OECD). Retrieved from. http://www.oecd.org/pisa/keyfindings/pisa-2012-results-overview.pdf

Patton, M. Q. (2002). *Qualitative research and evaluation methods*. Thousand Oaks: Sage.

Sadler, R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education, 5*(1), 77–84.

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

Urmston, A., Raquel, M., & Tsang, C. (2013). Diagnostic testing of Hong Kong tertiary students' English language proficiency: The development and validation of DELTA. *Hong Kong Journal of Applied Linguistics, 14*(2), 60–82.

Yorke, M. (2003). Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. *Higher Education, 45*(4), 477–501.

Part III
Addressing the Needs of Doctoral Students

Chapter 6
What Do Test-Takers Say? Test-Taker Feedback as Input for Quality Management of a Local Oral English Proficiency Test

Xun Yan, Suthathip Ploy Thirakunkovit, Nancy L. Kauper, and April Ginther

Abstract The Oral English Proficiency Test (OEPT) is a computer-administered, semi-direct test of oral English proficiency used to screen the oral English proficiency of prospective international teaching assistants (ITAs) at Purdue University. This paper reports on information gathered from the post-test questionnaire (PTQ), which is completed by all examinees who take the OEPT. PTQ data are used to monitor access to the OEPT orientation video and practice test, to evaluate examinee perceptions of OEPT characteristics and administration, and to identify any problems examinees may encounter during test administration. Responses to the PTQ are examined after each test administration (1) to ensure that no undue or unexpected difficulties are encountered by examinees and (2) to provide a basis for modifications to our administrative procedures when necessary. In this study, we analyzed 1440 responses to both closed-ended and open-ended questions of the PTQ from 1342 test-takers who took the OEPT between August 2009 and July 2012. Responses to these open-ended questions on the OEPT PTQ provided an opportunity to examine an unexpectedly wide variety of response categories. The analysis of the 3-year data set of open-ended items allowed us to better identify and evaluate the effectiveness of changes we had introduced to the test administration

X. Yan (✉)
Department of Linguistics, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, USA
e-mail: xunyan@illinois.edu

S.P. Thirakunkovit
English Department, Mahidol University, Bangkok, Thailand
e-mail: suthathip.thi@mahidol.ac.th

N.L. Kauper
Oral English Proficiency Program, Purdue University, West Lafayette, IN, USA
e-mail: nkauper@purdue.edu

A. Ginther
Department of English, Purdue University, West Lafayette, IN, USA
e-mail: aginther@purdue.edu

J. Read (ed.), *Post-admission Language Assessment of University Students*, English Language Education 6, DOI 10.1007/978-3-319-39192-2_6

process during that same period of time. Carefully considering these responses has contributed substantially to our quality control processes.

Keywords Test-taker feedback • Quality management • Speaking assessment • International teaching assistants • Semi-direct tests • Ethics and responsibility • Test administration

1 Introduction

Purdue University is a large, research-intensive US university located in Indiana, specializing in science and engineering, with a large and growing number of international students. Purdue's Oral English Proficiency Program (OEPP) was established in 1987 in response to the perceived crisis in higher education associated with the presence of, and dependence on, a growing number of international graduate students to teach undergraduate introductory courses. One reaction to the "foreign TA problem" (Bailey 1984) was to establish English language screening and training programs for prospective international teaching assistants (ITAs). At the time, many state governments were mandating screening and training programs (Oppenheim 1997), parents and students were bringing lawsuits against universities, and Purdue's program was established to protect the university from both. Today, ITA programs are well established, and ITA screening has become one of the most widely practiced forms of post-entry testing at large research universities in the United States.

From 1987 to 2001, to screen prospective ITAs the OEPP used the Speaking Proficiency English Assessment Kit (SPEAK), which is an institutional version of the Test of Spoken English developed by Educational Testing Service (ETS) (1985). In 2001, the program replaced the SPEAK with a locally developed test, the Oral English Proficiency Test (OEPT), which is semi-direct (using pre-recorded questions and no interlocutor) and computer-administered. The introduction of the OEPT resulted in a number of improvements in terms of construct representation and test administration, not the least of which was a reduction of two-thirds in the time required for test-taking and rating. Although computer-administered, semi-direct oral English testing is now widespread (notably in the internet-based TOEFL (iBT)), when we introduced the OEPT in 2001, test-taker preparation for and comfort with this testing format was less than assured.

In a situation in which an institution requires its students to take a test, every effort must be made to ensure that prospective examinees understand the motivation for the testing and have access to appropriate test preparation materials. The OEPP is primarily responsible for the provision of both test justification and preparation materials, but we share this responsibility with Purdue's Graduate School and with each of the more than 70 graduate departments and programs that require oral English screening. The post-test questionnaire (PTQ) was introduced with the

OEPT in 2001 in order to (1) track student access to and use of test preparation materials and (2) understand and monitor examinee perception of general OEPT characteristics. Section III of the PTQ, consisting of two open-ended questions, was added in 2009 in order to identify any problems that may have been missed in test-taker responses to the fixed-response items in Sections I and II. Monitoring examinee feedback through the PTQ has become a central component of our quality management process.

2 Literature Review

2.1 Test-Taker Feedback About Semi-direct Testing

Of particular interest in our context are studies examining test-taker feedback about semi-direct testing formats for oral proficiency testing. Given that the Speaking subsection of the TOEFL iBT is semi-direct and is taken by the majority of international applicants for North American universities to demonstrate required language proficiency, semi-direct oral proficiency testing can now be assumed largely familiar to prospective international examinees; however, familiarity does not ensure comfort with, or acceptance of, the procedures associated with the semi-direct format.

The benefits of semi-direct testing are largely associated with cost effectiveness and efficiency in that interviewers are not required and ratings of recorded performances can be captured, stored, and rated remotely after the real-time administration of the test. However, cost efficiency alone cannot justify the use of semi-direct formats, and researchers have considered the comparability of semi-direct and direct formats to determine whether examinees are ranked similarly across formats. In a comparison of the ACTFL Oral Proficiency Interview (OPI) to its semi-direct counterpart (the ACTFL SOPI), Stansfield and Kenyon (1992) reported a high degree of concurrent validity based on strong positive correlations (0.89–0.92) across direct and semi-direct formats. Shohamy (1994) also found strong positive correlations across a Hebrew OPI and SOPI but cautioned against assuming total fidelity of the formats as language samples produced in response to the direct OPI tended to be more informal and conversational in nature, while those produced in response to the SOPI displayed more formality and greater cohesion.

The absence of an interviewer can be seen as either a negative or positive attribute of the semi-direct format. The most obvious drawback in the use of semi-direct formats lies in the omission of responses to questions and in the lack of opportunity for responses to be extended through the use of interviewer-provided probes; that is, the apparent limitations to the validity of the format are due to the absence of interactivity. On the other hand, standardization of the test administration removes the variability associated with the skill and style of individual interviewers, resulting in an increase in reliability and fairness, in addition to cost effectiveness and efficiency.

Test-takers can reasonably be expected to value validity over reliability, and they may not appreciate cost effectiveness even if they benefit. Comparisons of semi-direct and direct formats typically include examinations of test-taker perceptions and historically these perceptions have favored direct oral testing formats over their semi-direct counterparts. McNamara (1987), Stansfield et al. (1990), Brown (1993), and Shohamy et al. (1993) report test-taker preferences for semi-direct formats ranging from a low of 4 % (Shohamy et al.) to 57 % (Brown); however, the preference for the semi-direct format reported by Brown seems to be an outlier. In a more recent comparison of test-taker oral testing preferences in Hong Kong, Qian's (2009) results suggest that a shift may be taking place. In his study, while 32 % of the respondents reported that they preferred the direct format, 40 % were neutral, so perhaps the strongly negative perceptions reported in former studies may be dissipating. Again, however, only 10 % of his respondents actually favored the semi-direct format. The remainder disliked both formats or had no opinion.

In a much larger scale study of test-taker feedback, Stricker and Attali (2010) report test-taker attitudes towards the TOEFL iBT after sampling iBT test-takers from China, Colombia, Egypt, and Germany. Stricker and Attali reported that, while test-taker attitudes towards the iBT tended to be moderately favorable overall, they were more positive towards the Listening and Writing subsections, and decidedly less favorable about the Speaking subsection. Test-takers were least favorable about the Speaking subsection in Germany, where 63 % disagreed that *The TOEFL gave me a good opportunity to demonstrate my ability to speak in English*, and in Colombia, where 45 % also disagreed with the statement. Although the researchers did not directly ask respondents to comment on direct versus semi-direct test formats, it seems safe to assume that the semi-direct format is implicated in their less favorable attitudes toward the Speaking subsection. Substantial percentages of respondents in Egypt (40 %) and China (28 %) also disagreed with the statement, but negative attitudes were less prevalent in these countries.

If a language testing program decides that the benefits of semi-direct oral testing are compelling, then the program should acknowledge, and then take appropriate steps to ameliorate, the negative test-taker attitudes towards the semi-direct format. Informing prospective test-takers about the use of the format and providing test preparation materials is an essential step in this process; however, the provision of information is only the first step. Collecting, monitoring, and evaluating test-taker feedback concerning their access to and use of test prep materials is also necessary.

2.2 *Quality Management Using Test-Taker Feedback*

Both under-researched and under-theorized, quality control of the administrative procedures associated with an operational test comprises the lion's share of the day-to-day work required to maintain a testing program. Saville (2012) is one of the few test developers who has addressed quality control in the testing literature. He defines quality management as "the planning and management of processes, which over time lead to improvements being implemented" (p. 399).

Test quality management processes and validation processes may intersect in the type of evidence collected. The two processes differ, however, in their purposes and in their use of evidence, quality management being a more practical process aimed at effecting improvements to testing processes rather than providing supporting warrants in a validity argument such as that articulated for TOEFL by Chapelle et al. (2008). But because quality management processes may result in improvements to test reliability and fairness, such procedures may affect, and be considered part of, the overall validation process.

Acknowledging the scarcity of literature on systematic quality management procedures for language testing, Saville (2012) emphasizes the importance of periodic test review for effective quality control (p. 408) and provides a quality management model that links periodic test review to the assessment cycle. He argues that quality control and validity go hand in hand. Saville's quality management process, like the test validation process, is iterative in nature and consists of the following five stages:

1. Definition stage: recognize goals and objectives of quality management;
2. Assessment stage: collect test-related feedback and identify potential areas for improvement;
3. Decision stage: decide on targeted changes and develop action plans accordingly;
4. Action stage: carry out action plans; and
5. Review stage: review progress and revise action plans as necessary. (pp. 408–409)

A comprehensive quality management process favors collection of multiple facets of information about a test, including information about and from test-takers. Shohamy (1982) was among the first language testers who called for incorporation of test-taker feedback to inform the test development process and test score use. Stricker et al. (2004) also advocated periodic monitoring of test-taker feedback. Although one might argue that feedback from test-takers can be very subjective, and some studies have shown that test-taker feedback is partially colored by proficiency level or performance (Bradshaw 1990; Iwashita and Elder 1997), there are certain types of information that only test-takers can provide. Only test-takers themselves can report on their access to and use of test preparation materials, their understanding of test instructions and materials, their experience of the physical test environment, and their attitudes about interactions with test administrators.

3 Research Questions

The purpose of the present study is to describe and discuss how test-taker feedback is used in our quality control procedures for the OEPT. We have found that test-taker feedback provides a practical starting point for improvement of the test and our administrative procedures. The following are questions/concerns that we address by examining responses to the three sections of the post-test questionnaire (PTQ).

1. To what extent are prospective test-takers provided information about the OEPT?
2. To what extent do they actually use OEPT test prep materials?
3. Do test-takers find selected characteristics of the test (prep time, item difficulty, representation of the ITA classroom context) acceptable and useful?
4. What experiences and difficulties do test-takers report?
5. Which aspects of the testing process may require improvement?
6. Do actions taken to improve test administration have an effect?

4 Context of the Study

4.1 The Oral English Proficiency Test (OEPT)

The OEPT is a local computer-based, semi-direct performance test, developed by the Oral English Proficiency Program (OEPP), to screen L2 English-speaking prospective teaching assistants for their oral English proficiency (See Ginther et al. 2010). The current version of the OEPT, in use since 2009, uses a six-point holistic scale of 35-40-45-50-55-60. A score of 50 or higher is required for ITA certification; that is, examinees who score 50, 55, 60 may be given a teaching assistantship without any restriction. Students who score 35, 40, or 45 may be required by their academic departments to enroll in an ESL communication course offered by the OEPP. Students pay no extra fees for the test or the course.

The OEPT is considered a mid-stakes test. A passing score on the test is one way for students to be certified in oral English proficiency before they can be employed as a classroom teaching assistant. Due to cuts in research funding in recent years, there are fewer research assistantships available and consequently an increase in competition for teaching assistantships. For this reason, the stakes of the OEPT test have increased for many graduate students who need teaching assistantships to fund their graduate studies.

The OEPT consists of twelve items with four different prompt types: text, graph, listening, and read aloud. (See Appendix A, for the OEPT item summary.) Test-takers are required to express opinions and speak extemporaneously on topics related to tasks associated with an academic setting. Each item allows 2 min for preparation and 2 min for response. Item analyses of the OEPT indicate that graph and listening tasks are the most difficult items (Oral English Proficiency Program 2013a, p. 29).

4.2 OEPT Score Use for Placement and Diagnostic Purposes

When examinees fail the test, their OEPT test record serves both placement and diagnostic functions for the ESL communication course for ITAs, a graduate-level course taught in the OEPP. OEPT scores allow the OEPP to group students into

course sections by proficiency level. Because the course is reserved exclusively for students who have failed the test, and course sections are small, instructors (who are also OEPT raters) use the test recordings to learn about their assigned students' language skills. Instructors listen to and rate test performances analytically, assigning scores to a student's skills in about a dozen areas related to intelligibility, fluency, lexis, grammar and listening comprehension, using the same numerical scale as the OEPT. By this means, and before the first class meeting, instructors begin to identify and select areas that a student will be asked to focus on improving during the semester-long course. Students then meet individually with instructors to listen to and discuss their test recording, and the pair formulate individual goals, descriptions of exercises and practice strategies, and statements of how progress towards the goals will be measured. This OEPT review process also helps students understand why they received their test score and were placed in the OEPP course. During subsequent weekly individual conferences, goals and associated descriptions on the OEPT review document may be adjusted until they are finalized on the midterm evaluation document. Scores assigned to language skills on the OEPT review become the baseline for scores assigned to those skills on the midterm and final course evaluations, providing one means of measuring progress. The same scale is also used for evaluating classroom assessment performances.

Not only is the OEPT test linked to the OEPP course by these placement, diagnostic and review practices, but test raters/course instructors benefit from their dual roles when rating tests and evaluating students. Course instructors must make recommendations for or against certification of oral English proficiency for each student at the end of the course. Training and practice as OEPT raters provide instructors with a mental model of the level of proficiency necessary for oral English certification; raters can compare student performances on classroom assessments to characteristics described in the OEPT scale rubrics and test performances. In turn, when rating tests, raters' understanding of the test scale and of the examinee population is enhanced by their experience with students in the course; OEPT examinees are not just disembodied voices that raters listen to, but can be imagined as students similar to those in their classes. Thus, the OEPT test and OEPP course are closely associated in ways both theoretical and practical.

4.3 The OEPT Practice Test Website

Two forms of the OEPT practice test, identical in format and similar in content to the actual test, are available online so that prospective test-takers may familiarize themselves with the computer-mediated, semi-direct format, task types, and content of the test. A description of the OEPT scale and sample item responses from actual test-takers who passed the test are also available to help test-takers understand the types and level of speaking and listening skills needed to pass the test. The practice test website also provides video orientations to university policies and the OEPP, and taped interviews with graduate students talking about student life.

4.4 Test Administration

The OEPT is administered in campus computer labs several times a month throughout the academic year. At the beginning of each testing session, test-takers are given an orientation to the test, in both written and oral, face-to-face formats, including directions for proper use of the headset and information about the test-user interface and the post-test questionnaire. Examinees are also given a brochure after taking the test which provides information about scoring, score use, and consequences of test scores.

4.5 Post-Test Questionnaire (PTQ)

Each examinee completes a PTQ immediately following the test. The questionnaire is part of the computer test program and generally requires 5–10 min to complete. The PTQ is divided into three sections. Section I consists of nine fixed-response items that elicit information about awareness and use of the OEPT Practice Test Website; Section II consists of 11 items that elicit information about the overall test experience and includes questions about the within-test tutorial, within-test preparation time and response time, and whether test-takers believe their responses to the OEPT actually reflect their ability. Each question in Part II uses either a binary scale (*yes* or *no*) or a 5-point Likert scale (*strongly agree*, *agree*, *no opinion*, *disagree*, or *strongly disagree*). Section III consists of the following two open-ended questions:

1. Did you encounter any difficulties while taking the test? Please explain.
2. We appreciate your comments. Is there anything else that you would like us to know?

All questionnaire responses are automatically uploaded to a secure database on the university computing system.

5 Method

We review OEPT survey responses after each test administration. In this paper, we will briefly discuss responses from Section I of the PTQ that were collected in the four test administrations during the fall semester of 2013 ($N = 365$), as these most accurately reflect our current state. The remainder of this paper will discuss our analysis of responses to the two open-ended questions collected over a 3-year period – 1440 responses from 1342 test-takers[1] who took the OEPT between August

[1] The number of responses is higher than the number of examinees because some examinees took the test more than once.

2009 and July 2012. All test-takers were matriculated graduate students from around the world, most in their 20s and 30s. The majority of test-takers came from the Colleges of Engineering (43 %) and Science (24 %). Responses to closed-ended questions from Part I of the survey were analyzed with Statistical Analysis System (SAS), version 9.3, in terms of frequency counts and percentages. Written responses to open-ended questions were coded into categories.

6 Results and Discussion

6.1 Responses to Closed-Ended Questions

6.1.1 Access to and Use of the OEPT Practice Test

Figure 6.1 presents item responses to questions about test-taker awareness and use of the OEPT Practice Test. The figure is interesting in several aspects. First of all, despite the fact that information about the practice test and the URL for the practice test is included in the admission letter that the Graduate School sends to each admitted student, slightly less than 60 % of our test-takers typically report that they were informed by the Graduate School about the OEPT. A higher percentage (87 %) report being informed by their departments about the OEPT requirement for prospective teaching assistants. A slightly lower percentage (80 %) report that they were advised to take the practice test, and 70 % report having completed the practice test.

We believe that 70 % completion of the practice test is problematic – especially when we take into account the negative perceptions of the semi-direct format reported in the literature. Furthermore, the absence of knowledge about the test format may negatively affect subsequent test-taker performance.

If we break on departments, we can see that prospective test-takers' completion of the practice test differs considerably across departments. In some departments, only 50 % of the examinees completed the practice test; in some smaller programs, no one completed the test, whereas in other departments, virtually all test-takers completed it (Fig. 6.2).

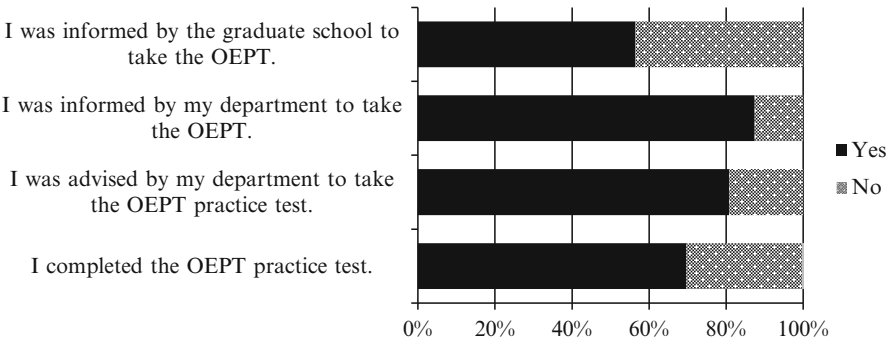


Fig. 6.1 Test-takers' awareness of the OEPT and the OEPT practice test ($N = 365$)

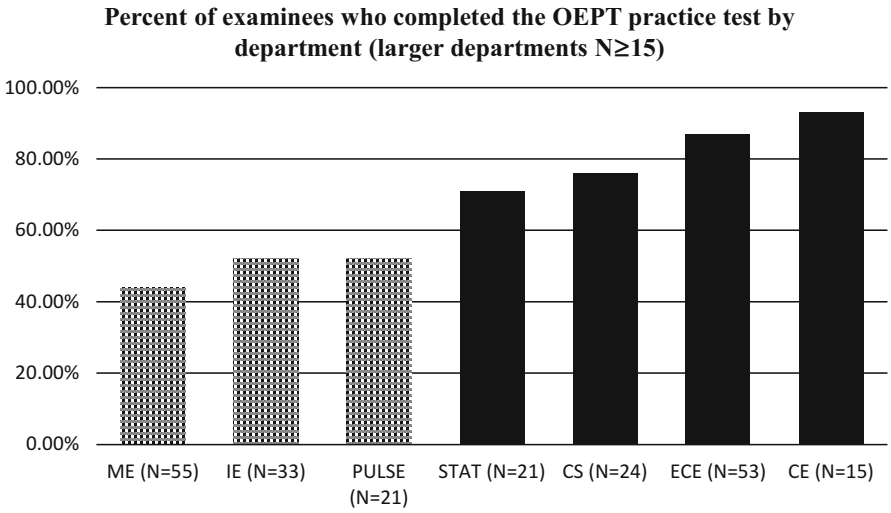


Fig. 6.2 Percent of examinees who completed the OEPT practice test (larger departments). *ME* Mechanical Engineering, *IE* Industrial Engineering, *PULSE* Purdue University Interdisciplinary Life Science, *STAT* Statistics, *CS* Computer Science, *ECE* Electrical Computer Engineering, and *CE* Civil Engineering

It is clearly the case that the provision of a practice test is not sufficient. Local testing programs must accept the added responsibility/obligation for tracking overall awareness and use by department, and then alerting departments about the relative success of their efforts to inform their students.

After tracking awareness and use since the OEPT became operational in 2001, we recently achieved our goal of 80 % completion of the practice test across all departments, and have set a new goal of 90 %. Our ability to hit 90 % is negatively influenced by turnover in departments of the staff/administrative positions that function as our liaisons. In addition, some test-takers will probably always decide to forego the opportunity.

6.1.2 Perceptions of the Quality of OEPT

Section II of the PTQ asks test-takers to provide information about the test itself. Responses to these items are presented in Fig. 6.3.

First, we ask whether test-takers consider themselves prepared to take the test. This item is somewhat problematic because respondents can interpret this question as referring to test prep or language proficiency. In either case, only 60 % of the test-takers over the 3-year period agreed or strongly agreed that they were prepared. Test-takers are more positive about test length and the instructions within the test. In both cases, at least 80 % agreed that both the test length and the instructions were acceptable.

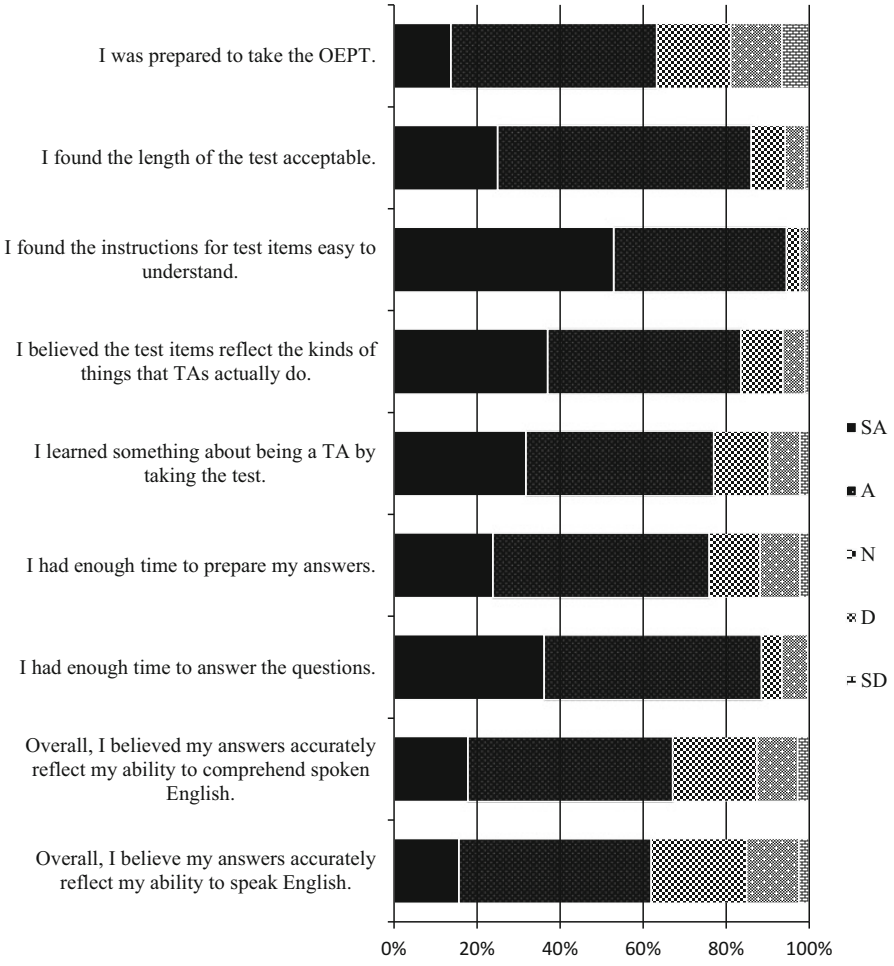


Fig. 6.3 Responses to items on the OEPT post-test questionnaire Part III ($N = 365$). *SA* Strongly agree, *A* agree, *N* No opinion, *D* Disagree, *SD* Strongly disagree

We ask the next two questions because we are always interested in understanding whether test-takers perceive that our items are reflective of actual TA tasks in classroom contexts and whether they have learned anything about being a TA by taking the OEPT. In both cases, around 80 % of the test-takers either agree or strongly agree that the items are reflective of TA tasks and that the test was informative about teaching contexts.

The next two questions ask whether test-takers had enough time (1) to prepare their responses and (2) to respond. When we developed the OEPT, we decided to allow test-takers 2 min to prepare and 2 min to respond to each item. Unlike other tests that allow only a very short prep time (e.g., TOEFL iBT speaking allows 15 s), we opted to turn down the pressure, perhaps at the expense of authenticity. Given

Table 6.1 Principal categories of test-taker comments on the OEPT PTQ part II

Category	Frequency	Percent
No difficulty	716	49.72
Positive	143	9.93
Difficulty	528	36.67
Miscellaneous	53	3.68
Total	**1440**	**100.00**

the computer-administered, semi-direct format, we believe that the provision of a relatively long prep time will perhaps ameliorate negative reactions to the computer administration and *bias for best* (Fox 2004). With regards to prep time, close to 80 % agree or strongly agree that 2 min is enough. More than 90 % agree that the 2 min response time is enough.

Finally, we ask whether test-takers believe that the test allows them to demonstrate their ability to comprehend and to speak English. Test-takers are least positive in response to these questions. For both, only slightly higher than 60 % either agree or strongly agree that their responses accurately reflect their ability to comprehend and speak English.

The next section will discuss test-taker responses to the two open-ended questions:

1. Did you encounter any difficulties while taking the test? Please explain.
2. We appreciate your comments. Is there anything else that you would like us to know?

6.2 Responses to Open-Ended Questions

In the initial stages of the coding process, we read and analyzed responses to open-ended questions several times and identified eight major categories of test-taker feedback: (1) Positive comments; (2) Comments that indicate no difficulties without further elaboration; Difficulties associated with: (3) test administration, (4) test preparation, (5) test design, (6) test-taker characteristics and performance, (7) other difficulties; and (8) Miscellaneous comments (See Appendix B for the coding scheme). Next, one coder coded all 1440 comments using the eight categories listed above. A second coder randomly coded 25 % of the comments. Inter-coder reliability, calculated as exact agreement, was 92.5 %. However, in our data, some comments on difficulties covered more than one topic. Therefore, these difficulty comments were coded multiple times according to the different topics mentioned (see Table 6.1); as a result, the total number of topics is higher than the total number of comments.

Responses to the open-ended questions showed that examinee experience with the OEPT over the 3-year period was overall positive. As seen in the last column of

Table 6.1, about 50 % of test-takers reported no difficulties or problems. In addition, about 10 % of test-takers wrote positive comments about their experience with the OEPT. Negative responses, i.e., comments reporting difficulties, which account for about 40 % of the total and cover a range of topics, were further classified into subcategories (see Table 6.2). These negative responses will be discussed in conjunction with actions and reactions on our part.

6.2.1 Positive Test-Taker Comments

Among positive comments, the most frequently mentioned topics were the test administration process, interactions with test administrators, authenticity and relevance of test content, and washback effects of the test, as illustrated by the following comments:

Table 6.2 Subcategories and individual topics of difficulties reported by test-takers

Subcategory	Frequency	Percent
Test design	**277**	**40.38**
Item preparation and response time	89	12.97
Difficulty of individual test items	89	12.97
Semi-direct test format	57	8.31
Test length	21	3.06
Test program interface	12	1.75
Authenticity of content	9	1.31
Test administration	**196**	**28.57**
Noise and other test environment	132	19.24
Equipment and supplies	51	7.43
Time and location of test	5	0.73
Test registration process	4	0.58
Test orientation process	4	0.58
Test preparation	**80**	**11.66**
Practice test and other test preparation materials	71	10.35
Awareness of the OEPT test	9	1.31
Test-taker characteristics and test performance	**69**	**10.06**
Test-taker physical/mental conditions	35	5.10
Concerns about examinees' limited language proficiency	34	4.96
Other difficulties	**64**	**9.33**
Total	**686**[a]	**100.00**[b]

[a]Total frequency of topics is higher than total number of difficulty comments in Table 6.1 because some comments covered more than one topic

[b]Total percentage may not add up to 100 due to rounding

The OEPT test was useful and the process of giving the test was reasonable and well arranged. (November 2011)

I find the topics presented here more relevant and engaging than the ones in TOEFL. (October 2010)

The test was administered very well. (August 2010)

The test was conducted in an excellent manner. (*August 2009*)

No difficulties encountered. System was nicely set up and program ran smoothly. (*August 2009*)

The test was not only a good indicator of preparedness for a TA position but also relevant in context including information that made me more aware of what it is to be a TA and a graduate student in general. Kudos to the OEPT team! (*January 2012*)

These comments reflect advantages of local language tests over large-scale language tests, many of which lie in the relative ease of anticipating and accommodating test-taker needs.

Another important advantage of local language tests has much to do with contextualization and positive washback effects. Even though large professional testing companies follow a set of rigorous procedures to monitor the quality of their language tests, the fact that these tests serve selection or certification purposes across a wide range of contexts somewhat limits the contextualization of test items and interpretations of test scores to a general level. The development of local language tests, on the other hand, is often dictated by particular purposes, such as placement of language learners in specified courses according to proficiency levels or certification of language proficiency for a specific purpose. These functionalities permit item writers to situate language tasks in contexts that represent the range of communicative tasks that may be found in the local context.

6.2.2 Negative Test-Taker Comments and Our Responses

Written comments indicating problems or difficulties fell into five broad categories, in descending order of frequency: test design, test administration, test preparation, test-taker characteristics and performance, and other difficulties (see the first column of Table 6.2 below). These broad categories were further broken down into individual topics, including—in descending order of frequency—test environment, item preparation and response time, difficulty of individual test items, online practice test and associated test preparation materials, semi-direct test format, testing equipment and supplies, and test-taker concerns about their language proficiency and test performance.

For purposes of test quality management, we are concerned with the appropriateness of test administration procedures, the availability of test preparation materials, and minimization of construct-irrelevant variance. While collection of test-taker feedback marked the starting point in this particular quality management process, determining and implementing possible improvements represented the more challenging part of the process. As we discovered, some test-taker comments referred to issues that test administrators and the OEPP had been making ongoing efforts to address, while other comments required no action plans.

6.2.3 Comments Linked to Ongoing Improvement Efforts

Among topics mentioned in the survey responses, we have been particularly attentive to two areas of difficulty that call for improvement efforts: noise in the test environment, and technical problems with the OEPT practice test website.

Noise in the Test Environment

Noise and distraction from other test-takers is the most commonly identified difficulty associated with test administration. OEPT testing labs are not equipped with sound-proof booths that separate computer stations, or with sound-proof headsets. As a result, survey comments often include complaints about noise, especially during the big testing week in August prior to the beginning of the academic year, when test sessions are generally full. Here are some sample comments on this topic:

I was able to listen to other students which distracted me sometimes. (*August 2009*)

The fellow test-takers talking beside each other was little disturbing. The headphone needs to be noise proof. (*August 2010*)

The students sitting around me talked too loud making me hard to concentrate on my test. (*August 2011*)

The noise problem has been an ongoing target of effort by test administrators since the OEPT began. Having exhausted possible and affordable technical solutions to the noise problem, we began to focus more narrowly on test administration. In 2010, we began to tell examinees during pre-test orientation that if they were speaking too loudly, we would request that they lower their volume a little by showing them a sign, being careful not to disturb them while preparing for or responding to an item. In 2011, we began to model for test-takers during the pre-test orientation the desired level of speaking volume. Most examinees have been able to comply with the volume guidelines, but there are occasionally one or two test-takers per test administration who have difficulty keeping their voice volume at a moderate level; the majority of those adjust, however, after being shown a sign that reads "Please speak a little more softly."

The decreasing numbers and percentages of complaints about noise during August test administrations in the three academic years covered in the data (Table 6.3) suggest that these noise-reducing efforts described above have been somewhat successful. However, despite some improvement, persistent noise issues require ongoing monitoring. In August 2012, we began to set up cardboard trifold presentation screens around each testing station. The screens do not reduce noise to a great degree, but they have been an overall improvement to the test-taking environment by reducing visual distractions to test-takers. The OEPP has also been involved recently in university-wide discussions about the need for a testing center on campus that could provide a more appropriate testing environment for the OEPT.

Table 6.3 Number of complaints about noise during August test administrations by year

Year	Frequency	Percent	Total responses
2009	21	9.29	226
2010	14	6.19	226
2011	9	3.77	239

Table 6.4 Number of complaints about accessibility to OEPT practice test website by academic year

Academic year	Frequency	Percent	Total response
2009–2010	30	6.40	469
2010–2011	11	2.24	492
2011–2012	5	1.04	479

Technical Problems with Practice Test Website

Parallel to concerted efforts to inform incoming graduate students about university rules pertaining to oral English proficiency testing, it has also been one of the OEPP's core interests to familiarize prospective OEPT examinees with the test item types and computer-based test format by means of the OEPT online practice test. It is, therefore, an important responsibility of test administrators to ensure accessibility to the online practice test and other test preparation materials.

Among test-taker comments in our data, the most frequently reported difficulties in terms of test preparation were technical problems associated with the practice test. Some test-takers reported not being able to access the practice test website or to download or open test files, as illustrated here:

> *Actually I had the trouble to download the practice test at the beginning. The web page always shows "error" and I don't know why. Finally I could download the test and it doesn't run well. During the practice test a label jumped out to show "no sound file". (November 2009)*
>
> *I did encounter a difficulty not while taking the test but while trying to take the practice test. I am a MAC OS user and I tried to follow the instructions for Linux/Mac that you gave in the download page of the OEPT Practice Test. Unfortunately after doing so, I still could not open the practice test. (August 2010)*

Despite efforts to keep the OEPT practice test website up-to-date and accessible, changing technology standards along with users across the globe attempting to access the site using a wide variety of hardware, software, and Internet access conditions resulted in persistent reports of website user problems from 2009 to 2012. Although Table 6.4 shows a trend of decrease in the numbers of complaints about accessibility to the practice test over the 3 years covered in our data, the OEPP considers even a small number of reported problems with use of the practice test website to be too many.

In response to ongoing test-taker comments about difficulties accessing the practice test, the OEPP contracted software developers in 2012 to replace the Java

Table 6.5 OEPT score distributions for examinees requesting more preparation and response time

	Time constraints			
	Preparation		Response	
Score	Frequency	Percent	Frequency	Percent
35	5	9.62	2	9.09
40	15	28.85	3	13.64
45	6	11.54	0	0.00
50	19	36.54	13	59.09
55	6	11.54	2	9.09
60	1	1.91	2	9.09
Total	**52**	**100.00**	**22**	**100.00**

version of the practice test with an internet-based version and to improve its compatibility with the Mac OS operating system. JAR files were dispensed with. To better track and evaluate technical problems, an online contact form was created for users to report problems. Monitoring of contact form submissions and of post-test questionnaire responses from 2012 to 2013 academic year test administrations have indicated very few user problems with the new version of OEPT practice test website since improvements were made.

6.2.4 Comments Eliciting No Immediate Action Plans

In addition to difficulties experienced by test-takers which prompted ongoing efforts to improve test conditions, three topics related to test design were identified by test-takers as posing difficulties: (1) timing for item preparation and response, (2) difficulty of particular test items, and (3) the semi-direct test format.

Preparation and Response Time

A small number of examinees ($n = 74$[2]) mentioned a desire or need for longer preparation and response time.

> *It would be better if the preparation time was longer. But it was acceptable*. (*August 2009*)
>
> *Very helpful if the prepation* [*sic*] *time can be longer would be nice*. (*January 2012*)
>
> *Nothing much except at one point I felt the time was little less for the response*. (*August 2010*)
>
> *Yes, I wish I had more time in some questions. In fact, I wish there was no timing at all. The timer either makes you talk faster or omit valious* [*sic*] *information*. (*August 2011*)

Table 6.5 captures a rather interesting phenomenon related to this topic: 19 complaints (almost 40 %) about preparation time and 13 complaints (almost 60 %) about

[2]This number is smaller than the frequency of the topic of item preparation and response time shown in Table 6.2 because some test takers requested both shorter item preparation and response time in their comments.

Table 6.6 OEPT score distributions for examinees commenting on the difficulty of test items

	Item type			
	Graph		Listening	
Score	Frequency	Percent	Frequency	Percent
35	1	3.23	5	8.62
40	8	25.81	19	32.76
45	11	35.48	16	27.59
50	9	29.03	14	24.14
55	2	6.45	3	5.17
60	0	0.00	1	1.72
Total	**31**	**100.00**	**58**	**100.00**

response time were made by examinees who passed the test with a score of 50 (fourth row). This observation may reflect examinees' attempts to ensure their best test performance, especially when some might have realized that their proficiency level was near the cut-off for passing.

Examinee efforts to maximize the quality of their test performance coincide with the original *bias for best* rationale of OEPT test developers for allotting 2 min of preparation time and 2 min of response time for test items. As mentioned above, to reduce test anxiety and elicit better test-taker performance, OEPT test developers decided to extend the length of preparation and response times to 2 min each. About 80 % of OEPT examinees in the PTQ Part I data agreed that 2 min was sufficient time for item preparation and item response.

Difficulty of Particular Test Items

A small number of examinees ($N=89$) commented on the difficulty of particular types of test items, predominantly graph and listening items. Perceived difficulty of these test items is illustrated by the following comments:

> *The information in the very beginning of those conversations and lectures which are only played once are difficult for me to get. Since it suddenly begins I hardly catch the information at the first few seconds. Thank you for asking.* (*January 2010*)
>
> *Some questions especially about bar chart and line chart are very difficult to understand because it is not my area of study. I was not very unfamiliar* [*sic*] *with the terms so it was hard to interpret or summarize.* (*April 2012*)

Moreover, as the results in Table 6.6 suggest, complaints about graph and listening items were observed from examinees across most score levels (second and fourth rows), suggesting that test-takers, regardless of their oral English proficiency level, tend to consider graph and listening items more difficult and challenging than text items. Compared with text items, graph and listening items are more integrated in terms of the cognitive skills required for processing the item information.

Statistical item analyses of the OEPT also identify graph as the most difficult, followed by listening items (Oral English Proficiency Program 2013a, p. 29). Therefore, test-taker perceptions of the greater difficulty of graph and listening items are legitimate. However, the OEPP relies partly on these integrated items to differentiate between higher and lower proficiency examinees. Because of their relatively higher difficulty levels, graph items—as suggested in the literature on using graphic items in language tests (Katz et al. 2004; Xi 2010)—are effective in eliciting higher-level language performances, e.g., the use of comparatives and superlatives to describe statistical values and differences and the use of other vocabulary needed to describe trends, changes and relationships between variables on a graph.

Nevertheless, it should be acknowledged that test-taker familiarity with graphs may have a facilitating effect on test performance (Xi 2005). Prospective OEPT examinees are encouraged to familiarize themselves with OEPT graph item formats by taking the OEPT practice tests. Familiarity with these item formats can also be increased by listening to the sample test responses and examining the OEPT rating scale on the practice test website. This information can help test-takers and test score users understand the intent of the graph items, which is to elicit a general description and interpretation of trends illustrated by the graph, rather than a recitation of all the numbers shown (as is also explained in the test item instructions). Yet, to address examinees' concerns from a customer service perspective, we continue to seek improved and multiple ways to stress to examinees the importance of taking the practice test and familiarizing themselves with the purpose and scoring of the test.

Semi-direct Test Format

Test-taker comments expressing dissatisfaction with the semi-direct format of the OEPT ($N=57$) appear to result mainly from two reported affective dispositions: unfamiliarity or discomfort with the computer and preference for direct interviews, as the comments below illustrate:

> *Facing to the computer without talking to person directly is a little hard for myself.* (*July 2010*)
>
> *Frankly speaking, I don't like this computer based test. It makes me feel nervous when I talk to a cool blood computer.* (*July 2011*)
>
> *I think that the test will be more objective if I was interviewed by a real person instead of recording the speeches in the computer.* (*October 2011*)

Although most literature on test-taker reactions suggests that test-taker computer skills do not exert a significant effect on performance on computer-based tests, some examinees' preference for an interview or human interaction versus recording to a computer is a valid concern. This concern could be addressed by efforts to communicate to examinees the rationale behind the choice of a computer-based test, in particular issues of reliability, standardization, and fairness.

Test-Taker Characteristics and Test Performance, Other Difficulties, and Miscellaneous Comments

In addition to the test-taker difficulties mentioned above, there were also some types of comments referring to conditions beyond the control or purview of test developers or test administrators. These comments include low-scoring examinees' concern about their language proficiency, test-takers' physical or mental conditions and test performance, and miscellaneous requests and suggestions. Some sample comments of these types are provided below:

No problem test went smoothly except I was feeling cold and had sore throat. (*August 2009*)

May be drining [*sic*] *water fecility* [*sic*] *should be provided. Throat becomes dry sometimes while speaking continuously*. (*August 2009*)

Nothing in particular but some snacks after the test would be nice as a reward for finishing the test. (*November 2010*)

Sorry about evaluating my record maybe it is the worst English you ever heard. (*August 2010*)

Sometime [*sic*] *I just forgot the words I want to use when I am speaking*. (*August 2009*)

I was very nervous so there are some cases where I paused and repeated one thing a number of times. (*August 2011*)

The addition in 2009 of Part III to the post-test questionnaire (the two open-ended questions) was made to open up a channel of communication for test-takers to express freely their reactions to the test. This channel is as much an effort to involve test-takers—an important group of stakeholders—in the quality management procedures as an assurance to the test-takers that their voices are heard by the language testers. Reading comments such as those above contribute to test administrators' awareness of examinees as individuals, each with unique feelings, dispositions, and personal circumstances. This awareness is in keeping with the recognition of the humanity of test-takers, mentioned in the International Language Testing Association (ILTA) Code of Ethics (2000).

One consequence resulting from these types of comments was the creation of a new test preparation brochure that was distributed on paper in 2013 to all graduate departments and added electronically to the OEPP website (Oral English Proficiency Program 2013b). In addition to providing general information about the purpose of the test, the brochure directs readers to the practice test website, advises students to bring a bottle of water and a jacket or sweater with them to the test, and alerts prospective examinees about issues of noise in the test environment. The purpose of the brochure is to facilitate test registration and administration, but as with most attempts to better inform constituents, the challenge is to get the brochure in the hands (literally or figuratively) of prospective examinees and for them to read and understand it.

7 Conclusion

In this study, we examined 3 years of test-taker feedback as part of a quality management process for the OEPT. Although there had been earlier periodic reviews of responses to the OEPT PTQ Sections I and II in addition to monthly reviews of

test-taker comments on the PTQ Section III, a systematic review of test-taker comments collected from a longer period of time has offered us substantial benefits not possible with smaller data sets from shorter time spans. The process was enlightening in that it offered a different perspective of the data, allowing us to observe trends over time and subsequently to better identify and evaluate possible changes or improvements to the test. The open-ended questions on the OEPT PTQ in particular have provided an opportunity to collect a wide variety of information that contributes to the quality management process. We can therefore recommend practices of this sort to local testing organizations that use similar instruments for collecting test-taker feedback but do not have a large number of examinees on a monthly basis.

While not all feedback requires action, all feedback should be reviewed and considered in some way. Test developers must examine test-taker feedback in relation to the purpose of the test and the rationale behind the test design. Although an important purpose of quality management is to minimize construct-irrelevant variance, test developers must hold a more or less realistic perspective of what they can do given the parameters of their testing contexts.

The choice to collect and examine test-taker feedback as a routine practice stems not only from recommendations to involve test-takers in quality management procedures as a best practice for local language testing organizations, but also from our recognition of test-takers as important stakeholders in the OEPT. In regard to quality control, the European Association for Language Testing and Assessment (EALTA) Guidelines for Good Practice (2006) mention that there needs to be a means for test-takers to make complaints. The ILTA Code of Ethics (2000) states that the provision of ways for test-takers to inform language testers about their concerns is not just their right but also a responsibility. However, in order for test-takers to act on that responsibility, they must be given opportunities to do so. Having a channel built in to the test administration process is a responsible way for language testers to facilitate the involvement of test-takers in quality management.

Appendices

Appendix A: OEPT2 Item Summary

Item	Title	Type	Expected response
1	Area of study	Text	Describe your area of study for an audience of people not in your field
2	Newspaper headline	Text	Given an issue concerning university education, express an opinion and build an argument to support it
3	Compare and Contrast	Text	Based on 2 sets of given information, make a choice and explain why you made it
4	Pros and Cons	Text	Consider a TA workplace issue, decide on a course of action, and discuss the possible consequences of that action

(continued)

Item	Title	Type	Expected response
5	Respond to complaint	Text	Give advice to an undergraduate concerning a course or classroom issue
6	Bar chart	Graph	Describe and interpret numerically-based, university-related data
7	Line graph	Graph	Describe and interpret numerically-based, university-related data
8	Telephone message	Listening	Relay a telephone message in a voicemail to a peer
9	Conversation	Listening	Summarize a conversation between a student and professor
10	Short lecture	Listening	Summarize a lecture on a topic concerning graduate study
11	Read aloud 1	Text	Read aloud a short text containing all the major consonant and vowel sounds of English
12	Read aloud 2	Text	Read aloud a passage from a University policy statement containing complex, dense text

Appendix B: Coding Scheme for Responses to Open-Ended Questions

Category	Subcategory	Topics	Codes
1. Test administration	1. Test environment	1. Noise	1-1-1
		2. Room temperature	1-1-2
	2. Equipment and supplies	1. Paper and pencil supplies	1-2-1
		2. Headset	1-2-2
	3. Test registration process	–	1-3
	4. Test orientation process	–	1-4
	5. Time and location of test administration	–	1-5
2. Test preparation	1. Online practice test	1. Practice test	2-1-1
		2. Sample responses	2-1-2
	2. Awareness of the OEPT test	–	2-2

(continued)

Category	Subcategory	Topics	Codes
3. Test design	1. Item preparation and response time	1. Insufficient preparation time	3-1-1
		2. Insufficient response time	3-1-2
		3. Too much preparation time	3-1-3
		4. Too much response time	3-1-4
	2. Test length	5. The test is too long	3-2-1
		6. The test is too short.	3-2-2
	3. Difficulty of individual test items	1. Graphic items	3-3-1
		2. Listening items	3-3-2
	4. Test program interface	–	3-4
	5. Semi-direct test format	–	3-5
	6. Authenticity of content	–	3-6
4. Test-taker characteristics and test performance	1. Test-taker physical/mental conditions	–	4-1
	2. Concerns about examinees' limited language proficiency	–	4-2
5. No problems indicated	–	–	5
6. Positive comments	1. Test administration	–	6-1
	2. Test design	–	6-2
	3. The practice test	–	6-3
7. Miscellaneous	–	–	7

References

Bailey, K. (1984). The foreign ITA problem. In K. Bailey, F. Pialorsi, & J. Zukowski-Faust (Eds.), *Foreign teaching assistants in U.S. universities* (pp. 3–15). Washington, DC: National Association for Foreign Affairs.

Bradshaw, J. (1990). Test-takers' reactions to a placement test. *Language Testing, 7*(1), 13–30.

Brown, A. (1993). The role of test-taker feedback in the test development process: Test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing, 10*(3), 277–301.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.

ETS (Educational Testing Service). (1985). *A guide to SPEAK*. Princeton: Educational Testing Service.

European Association for Language Testing and Assessment. (2006). *Guidelines for good practice in language testing and assessment.* Retrieved August 1, 2013, from http://www.ealta.eu.org/documents/archive/guidelines/English.pdf

Fox, J. (2004). Biasing for the best in language testing and learning: An interview with Merrill Swain. *Language Assessment Quarterly, 1*(4), 235–251.

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing, 27*(3), 379–399.

International Language Testing Association. (2000). *Guidelines for practice*. Retrieved August 1, 2013, from http://www.iltaonline.com/index.php?option=com_content&view=article&id=57&Itemid=47

International Language Testing Association. (2010). *Guidelines for practice*. Retrieved August 1, 2013, from http://www.iltaonline.com/index.php?option=com_content&view=article&id=122&Itemid=133

Iwashita, N., & Elder, C. (1997). Expert feedback? Assessing the role of test-taker reactions to a proficiency test for teachers of Japanese. *Melbourne Papers in Language Testing, 6*(1), 53–67.

Katz, I. R., Xi, X., Kim, H., & Cheng, P. C.-H. (2004). *Elicited speech from graph items on the Test of Spoken English* (TOEFL research report. No. 74). Princeton: Educational Testing Service.

McNamara, T.F. (1987). Assessing the language proficiency of health professionals. Recommendations for the reform of the Occupational English Test (Report submitted to the Council of Overseas Professional Qualifications.) Department of Russian and Language Studies, University of Melbourne, Melbourne, Australia.

Oppenheim, N. (1997). How international teaching assistant programs can protect themselves from lawsuits. (ERIC Document Reproduction Service No. ED408886).

Oral English Proficiency Program. (2013a). *OEPT technical manual*. West Lafayette: Purdue University.

Oral English Proficiency Program. (2013b). *Preparing for the oral English proficiency test: A guide for students and their Departments*. West Lafayette: Purdue University.

Qian, D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly, 6*, 113–125.

Saville, N. (2012). Quality management in test production and administration. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 395–412). New York: Routledge Taylor & Francis Group.

Shohamy, E. (1982). Affective considerations in language testing. *Modern Language Journal, 66*(1), 13–17.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing, 11*(2), 99–123.

Shohamy, E., Donitsa-Schmidt, S., & Waizer, R. (1993). *The effect of the elicitation mode on the language samples obtained in oral tests.* Paper presented at the 15th Language Testing Research Colloquium, Cambridge, UK.

Stansfield, C. W., & Kenyon, D. M. (1992). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System, 20*(3), 347–364.

Stansfield, C. W., Kenyon, D. M., Paiva, R., Doyle, F., Ulsh, I., & Cowles, M. A. (1990). The development and validation of the Portuguese speaking test. *Hispania, 73*, 641–651.

Stricker, L. J., & Attali, Y. (2010). *Test takers' attitudes about the TOEFL iBT* (ETS Research report. No. RR-10-02). Princeton: Educational Testing Service.

Stricker, L. J., Wilder, G. Z., & Rock, D. A. (2004). Attitudes about the computer-based test of English as a foreign language. *Computers in Human Behavior, 20*, 37–54.

Xi, X. (2005). Do visual chunks and planning impact the overall quality of oral descriptions of graphs? *Language Testing, 22*(4), 463–508.

Xi, X. (2010). Aspects of performance on line graph description tasks: Influenced by graph familiarity and different task features. *Language Testing, 27*(1), 73–100.

Zeidner, M., & Bensoussan, M. (1988). College students' attitudes towards written versus oral tests of English as a foreign language. *Language Testing, 5*(1), 100–114.

Chapter 7
Extending Post-Entry Assessment to the Doctoral Level: New Challenges and Opportunities

John Read and Janet von Randow

Abstract Since 2011, first-year doctoral candidates at The University of Auckland have been required to take the Diagnostic English Language Needs Assessment (DELNA) as part of a university-wide programme to identify students who are in need of significant language enrichment. This step was taken in response to research by the School of Graduate Studies suggesting that language difficulties often impacted on effective relationships between supervisors and their doctoral candidates, and progress in general. To ensure that such difficulties are addressed early in their study, candidates whose performance in DELNA indicates they need further language development attend an advisory session to discuss appropriate academic English enrichment programmes, and set specific goals to be achieved by the end of their provisional year. Although the doctoral learning process has often been studied from a variety of perspectives, there has been considerably less investigation of the language learning experiences of international doctoral students. This study focuses on 20 doctoral candidates in their first year and includes their reaction to DELNA, their response to the language advice they received, their evaluation of the language enrichment activities they engaged in, the strategies they used to adapt to their new environment, and their relationship with their supervisors. Results suggest that they welcome the fact that the University is proactive in responding to their language needs, and find that the specific programmes undertaken have increased their confidence in writing and their ability to express themselves better in all academic contexts.

J. Read (✉)
School of Cultures, Languages and Linguistics, University of Auckland, Auckland, New Zealand
e-mail: ja.read@auckland.ac.nz

J. von Randow
Diagnostic English Language Needs Assessment, University of Auckland, Auckland, New Zealand
e-mail: janetvonrandow@gmail.com

J. Read (ed.), *Post-admission Language Assessment of University Students*, English Language Education 6, DOI 10.1007/978-3-319-39192-2_7

Keywords Doctoral students • Diagnostic English Language Needs Assessment (DELNA) • Academic language development • Language advising • Conversation groups • Doctoral supervisors

1 Introduction

The number of international doctoral candidates enrolling in New Zealand universities has grown substantially over the past decade and shows every indication of continuing to do so. At the University of Auckland in 2013, 50 % (352) of new doctoral candidates were international students, and students from overseas now make up 40 % of the whole doctoral cohort. The majority of these students have English as an additional language (EAL) and, while all face a language challenge initially as they adapt to the new environment, many need considerable assistance to manage the language demands of their doctoral programmes. According to the University's Graduate School, this has added a certain amount of stress to the supervisory relationship (Lutz-Spalinger 2010) and, as the University intends to increase the number of doctoral completions to 500 a year in the next 7 years (The University of Auckland 2013a), the pressures are likely to increase.

Apart from the difficulties that all international students encounter, doctoral candidates face particular challenges in their new social and academic environment. Often they are mature students who need to settle their families in suitable accommodation and arrange schooling for their children once they arrive. Psychologically, they may experience a loss of status as they adjust to a kind of apprentice role after having been well-regarded academics or researchers in their home countries (Cotterall 2011; Fotovatian 2012). For doctoral students with English as an additional language, linguistic challenges which are well attested in the literature can impose an additional burden (Braine 2002; Nagata 1999; Strauss 2012). Whereas universities in Australia and New Zealand have made considerable efforts to address the language needs of their *undergraduate* students, it would seem that addressing those of doctoral candidates has not been a priority in most institutions (Benzie 2010).

In order to tackle this issue, the School of Graduate Studies at the University of Auckland introduced a policy in 2011 that all incoming doctoral students should take a post-admission English assessment and should be obliged to act on the advice subsequently given to enhance their academic language skills, where necessary. The policy applies to both international and domestic students, regardless of their language background. In this paper, then, we outline the background to the introduction of the policy and then report on a study which investigated the experiences and perceptions of a group of international doctoral candidates who undertook the assessment and follow-up language enhancement activities during the first 2 years of implementing the policy. The research can be seen as providing evidence for the validity of the assessment, in terms of its impact on the students' development of their academic language ability.

2 Review of the Literature

Much of the literature on international doctoral candidates in English-speaking countries focuses on their communicative needs. As they pursue their studies, these students must improve their language proficiency and build relationships within their departments and the wider university, with their peers, their supervisors and other significant people on the campus (Braine 2002). Building these relationships and adapting to the new academic environment is more difficult when the student has somewhat limited proficiency in English (Sawir et al. 2012). Many international EAL students, having achieved the minimum IELTS or TOEFL score required for admission, do not expect to encounter significant language-related difficulties (O'Loughlin 2008), while others believe their communication skills will improve simply by being in an English-speaking environment (von Randow 2013). However, opportunities to interact in English with people in the university, and thus to develop their communication skills, are often lacking (Benzie 2010; Cotterall 2011; Seloni 2012). This means that a vicious cycle is created as embarrassment caused by lack of practice impedes progress (Manathunga 2007b).

As Cotterall (2011, pp. 184–85) notes, one distinctive feature of Australian (and New Zealand) doctoral degrees, following the British tradition, is that there is normally no significant coursework component. This contrasts with the typical North American doctoral programme in which candidates take a whole range of courses, offering them multiple opportunities to interact with both professors and fellow students in a classroom setting over several semesters. Australasian university departments vary greatly in the extent to which they compensate for the absence of courses by creating a research community through which doctoral candidates can participate in seminars, project teams or informal social events. Thus, international EAL students may find themselves isolated from English speakers and critically dependent on their supervisors for guidance and support.

The relationship between supervisor and doctoral candidate has been extensively researched (Aitchison et al. 2012; Benzie 2010; Cotterall 2011; Grant 2003; Manathunga 2007a), often from a somewhat negative standpoint which emphasizes the difficulties experienced by both parties (Edwards 2006; Strauss 2012). When one or both have English as an additional language, it has been shown that even greater difficulties are likely to arise (Knight 1999; Nagata 1999). In this situation, according to Cotterall (2011), supervisors need to bear in mind that their EAL students are "diversely talented, multiliterate, culturally sophisticated individuals" (p. 49). The stereotyping or 'homogenizing' of these students, not just by staff (Sawir et al. 2012; Knight 1999; Fotovatian 2012) but also by their peers (Benzie 2010), has been shown to inhibit the acquisition of academic language skills. One study (Strauss 2012) found that supervisors who were struggling to interact with their EAL doctoral candidates become frustrated and impatient, resulting in both parties suffering. When giving feedback on written work, there is a tendency for supervisors to focus on the surface-level language issues (Basturkmen et al. 2014)

and, often in the natural sciences, they feel that they do not have the expertise to tackle language issues (Aitchison et al. 2012; Murray 2012).

EAL doctoral candidates with low levels of language proficiency will struggle to acquire the discipline-specific academic literacy which is essential for graduate studies (Braine 2002), especially considering that at the end of 3–4 years they have to produce a written thesis (Owens 2006). Assisting such students with appropriate English language enrichment as early as possible in their candidature has been shown to be essential if students are to rise to the language challenge (Manathunga 2014). Other researchers in the field of doctoral study support this view, suggesting that such students should have an assessment on arrival and ongoing support as required (Sawir et al. 2012), that language "should be taken seriously … and should be addressed very early in the candidature" (Owens 2007, p. 148) and that language enrichment be ongoing for students "who require support in refining their use of English over time" (Manathunga 2014, p. 73). Addressing language needs early should give the EAL doctoral candidates the confidence to engage more actively with their peers, their department and most importantly their supervisors (Benzie 2010; Sawir et al. 2012), thus positively influencing their learning (Seloni 2012).

Early identification of language needs in a systematic way seems to call for a post-admission assessment programme. Although numerous Australian universities have introduced post-entry language assessments (PELA) at the undergraduate level (see Dunworth 2009; Dunworth et al. 2013; Read 2015, Chap. 2), it appears that only the University of Auckland in New Zealand has expanded the scope of its assessment to cover doctoral candidates as well. The University's Diagnostic English Language Needs Assessment (DELNA) is described in some detail below.

One feature which distinguishes a PELA from the kind of placement test that international students commonly take when they enter an English-medium university is the scheduling of an individual advisory session after the students have completed the assessment, to discuss their results and recommend options for enhancing their language development. These sessions have been shown to work well with undergraduate students in Canada (Fox 2008) and Australia (Knoch 2012). While some authors argue that the individual consultation is not cost-effective (Arkoudis et al. 2012), it would seem that, when doctoral students may already be somewhat demoralised by the language difficulties they have been experiencing, a one-on-one meeting to follow up the assessment is highly desirable (Aitchison 2014; Laurs 2014). This session enables them to discuss their unique language needs and be listened to. It also provides information about specific language resources and initiates the networking that Carter and Laurs (2014) have described as important for doctoral candidates as they embark on their studies.

Any evaluation of a language support programme of this kind needs to focus on its impact in terms of positive outcomes for the students. Thus, Knoch and Elder (2013; this volume, Chap. 1) have developed a framework for the validation of post-entry assessments which gives prominence to the consequences of the assessment. As these authors argue, "the consequences of using the PELA and the decisions informed by the PELA should be beneficial to all stakeholders" (Knoch and Elder 2013, p. 60). The stakeholders here include university faculties and departments,

supervisors and, of course, the students. In the first instance, "the success of any PELA initiative relies on uptake of the advice stemming from test results" (2013, pp. 52–53). In his recent book, Read (2015, Chap. 10) has presented the available evidence for the impact of DELNA on the academic language development of undergraduate students at Auckland. The present study represents an initial investigation of the consequences for student learning of extending the DELNA requirement to doctoral students. It is necessary, then, to give some background information about how DELNA operates and what provisions are made by the University to enhance the academic language skills of doctoral candidates.

3 Assessing Doctoral Students at Auckland

3.1 Background

The University of Auckland introduced a post-entry language assessment, the Diagnostic English Language Needs Assessment (DELNA), for incoming undergraduate students from 2002 (DELNA is described further below). During the period until 2011, a total of 66 international doctoral candidates were also required to take DELNA as a condition of admission to the university. These were mainly students who had been accepted into a doctoral programme on the basis of an excellent academic record in their own country, even though they did not meet the University's English language requirement for international students, defined in terms of a minimum score on one of the major proficiency tests (www.auckland.ac.nz/en/for/international-students/is-entry-requirements/is-english-language-requirements.html). The students had, therefore, been granted a "language waiver", on condition that they completed the DELNA Diagnosis and engaged in a recommended language enrichment programme, if required. However, there was no effective procedure to ensure they complied with this condition and, up to 2011, only 27 of the 45 students whose assessment results showed that they should undertake language enrichment actually took the appropriate follow-up action.

As language waivers became more common and the overall numbers of EAL doctoral candidates increased, many students and their supervisors encountered difficulties in communicating effectively with each other. This became a major concern for the School of Graduate Studies and, after researching the issue, the School staff reported to their Board that poor communication skills were having a negative impact:

> There are too many examples of supervisory relationships foundering due at least in part to a lack of mutual understanding. The supervisor fixates on poor language skills; the student doesn't know what is wanted; frustration and resentment mount until they can no longer work together. (Lutz-Spalinger 2010, p. 2)

The response of the Board of Graduate Studies was to make the DELNA assessment a requirement for all incoming doctoral candidates, including domestic students

with English as a first language. In 2011, therefore, completion of the DELNA process became one of the goals set for doctoral candidates in their first year of registration, which is known as the provisional year (The University of Auckland 2013b). Once the provisional year goals are achieved, the registration is confirmed. As a result, since 2011 more than 400 doctoral candidates who were required to complete the DELNA Diagnosis and take up a language enrichment programme have complied with this provision.

3.2 The DELNA Process

As originally designed for use with undergraduate students, DELNA is a two-tiered assessment. First there is a 30-min computer-based Screening, which comprises an academic vocabulary task and a timed reading (cloze-elide) task and is used to exempt students above a certain cut score from further assessment (Elder and von Randow 2008). Those whose Screening scores are below the cut score proceed to the Diagnosis, a 2-h pen and paper assessment of listening, reading and writing. The results of the Diagnosis are reported in DELNA Bands on a 6 point scale from 4 to 9:

- Bands 4 and 5: students are at risk of failing their university courses
- Band 6: further language instruction is needed
- Band 7: students should work on their language skills independently
- Bands 8 and 9: no further language enrichment is needed

Students who receive an average band of 6.5 or below are informed that they should make an appointment with a DELNA Language Adviser to discuss their results and receive advice on an appropriate language enrichment programme.

The same broad procedure has been adopted for doctoral candidates, with some minor modifications. Those who receive a language waiver proceed directly to the Diagnosis phase, as they did before the new policy was introduced. Otherwise, the doctoral students take the Screening first, and typically 60 % of them are exempted on that basis. The others go on to the Diagnosis, which has the same listening and reading tasks as for undergraduate students but, given the importance of writing at the doctoral level, they complete an extended reading-writing task which requires them to draw on source information from two short input texts.

Students whose average DELNA band is below 7 on the Diagnosis must then attend a language advisory session in which the adviser discusses with them in some detail their strengths and weaknesses, particularly in writing. The students work with the adviser to create an individual programme of language enrichment to develop their English language skills during their provisional year and beyond. The resulting advisory report is sent to the student's doctoral supervisor, with copies to the postgraduate adviser for the student's faculty, the University's English Language Enrichment (ELE) centre and to the Graduate School. The student also receives a form on which they record all the enrichment activities they undertake during the

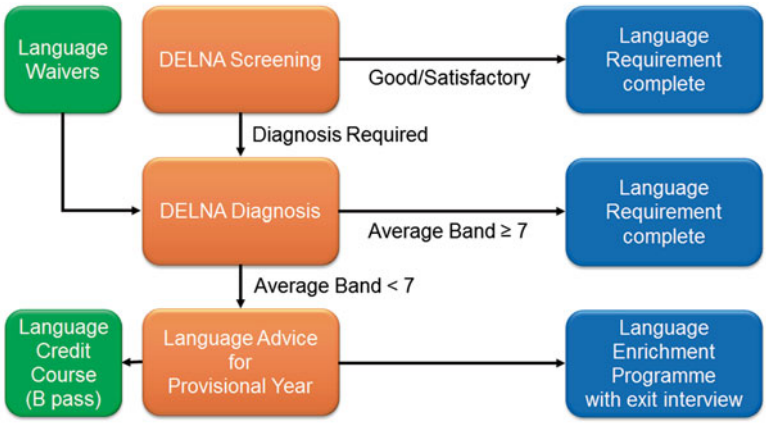


Fig. 7.1 The DELNA process for doctoral candidates

year, to form the basis of an exit interview with the doctoral consultant at ELE. The whole process is outlined in Fig. 7.1.

There are two main forms of language enrichment available at the University.

- The first consists of credit courses in academic English, English writing and scientific communication. Taking one of these courses, as appropriate to their needs, is a requirement for students with a language waiver and those whose band scores in the Diagnosis are below 6. No additional tuition is charged but the students must complete the course requirements with a minimum grade of B. For more advanced candidates, there is a postgraduate credit course on Developing Academic Literacy, which is of particular value for Arts students.
- The other main type of enrichment consists of activities which involve more individual motivation and effort on the student's part. The English Language Enrichment (ELE) centre offers individual consultations with an adviser, online and print resources, and weekly small-group conversation sessions – all designed specifically for EAL students (www.library.auckland.ac.nz/ele). ELE is a specialist unit within Student Learning Services (SLS), which runs workshops and provides resources for doctoral candidates on topics such as thesis writing, academic integrity, accessing the literature and presenting research (www.library.auckland.ac.nz/study-skills/postgraduate). The adviser also informs the candidate about individual learning opportunities such as accessing academic English podcasts.

There is an ongoing evaluation procedure for DELNA which is well established in the case of undergraduate students and has now been extended to doctoral candidates as well. Students who have taken the DELNA Diagnosis receive an email at the end of their first semester of study, requesting that they complete an anonymous online questionnaire about their experience of DELNA and their first semester of study. Over the last 11 years, the response rate has been about 22 %. Those who submit the

questionnaire are then invited to provide their contact details if they are willing to participate in a follow-up interview with a DELNA staff member to express their views and experiences at greater length. Typically three to four interviews have been conducted each semester. The interviews are quite separate from the process of determining whether the students have fulfilled the post-DELNA requirement to enhance their academic English skills at the end of their provisional year.

4 The Study

Thus, drawing on both the literature on doctoral students and the experience to date with DELNA as a programme for undergraduate students, the present study aimed to obtain some preliminary evidence of the effectiveness of the DELNA requirement and follow-up action in enhancing the academic language development of doctoral students at Auckland. More specifically, the following research questions were addressed:

1. How did the participants react to being required to take DELNA as they began their provisional year of registration?
2. How did they respond to the advisory session and advice?
3. How did they evaluate the language enrichment activities they were required or recommended to undertake?
4. What other strategies did they adopt to help them cope with the language demands of their provisional year?
5. What role did the doctoral supervisors play in encouraging language development?

4.1 Participants

The participants in this study were 20 doctoral students who took the DELNA Diagnosis between 2011 and 2013 and who agreed to participate in the evaluation interview. As explained above, they were very much a self-selected group and thus a non-probability sample. A demographic profile of the participants is presented in Table 7.1. In an informal sense, the group was representative of the international student cohort at the doctoral level in having Engineering as the dominant faculty and Chinese as the most common first language. Seven of the students had language waivers, which meant that they had not achieved the minimum IELTS or TOEFL scores before admission, whereas the other 13 had not reached the required cut score in the DELNA Screening. The four students with a band score of 5 from the DELNA Diagnosis were at particular risk of language-related difficulties, but the 13 with a score of 6 were also deemed in considerable need of language enrichment.

Table 7.1 The study participants

Gender		**Faculty**	
Male	12	Engineering	11
Female	8	Business School	3
First language		Science	3
Chinese	10	Education	2
Farsi	5	Creative arts and industries	1
Indonesian	2	**Language waiver**	
Sinhala	1	Yes	7
Thai	1	No	13
Vietnamese	1	**Overall DELNA band score**	
Years in New Zealand		5	4
1 or less	12	6	13
2	4	7	3
3	2		
4	2		

4.2 Data Sources

The main source of data for this study was the set of interviews with the 20 doctoral candidates. The interviews were transcribed and imported into NVivo, where they were double-coded line by line to extract the salient themes for analysis. Other data sources available to the researchers were: the candidates' results profiles generated from the DELNA Diagnosis, their responses to the online evaluation questionnaire, and their exit reports on the language enrichment undertaken by the end of the provisional year. However, the analysis in this chapter focuses primarily on the interview data.

4.3 Results

The findings will be presented in terms of responses to the research questions, drawing selectively on the data sources.

4.3.1 How Did the Participants React to Being Required to Take DELNA as They Began Their Provisional Year of Registration?

In the online questionnaire 19 of the participants were "happy to do DELNA" and found it a "fair assessment of their language ability"; just one remained neutral on both counts. In the interviews later, they elaborated on these initial judgements and 13 still agreed that they had no problem being asked to do DELNA. One commented

that "IELTS got you in, but that was only the beginning", and another advised others to do it because "it could record your English level and … improve your English".

I think it's okay because you know they might have some specific goals to see the English level of the students. I think it's okay. (#4 Engineering student, China)

Seven expressed certain concerns: two were nervous beforehand about taking DELNA; one found that post-DELNA language study interrupted his research project; two noted that their supervisors suddenly focused more on their language skills; and one was worried she would be sent home if her English skills were found wanting. Another felt that she had passed IELTS and so her "English was over", a perception that is not uncommon.

Actually after I know I need to take DELNA I am a little worried – other English examination – I passed the IELTS already you know – my English is over. (#7 Engineering student, China)

I think, oh my God, another exam, I hope I can pass! I felt like that, I thought I just worry about if I could not pass. (#8, Engineering student, China)

Before I take the DELNA I didn't want to do it, to be honest. I am very afraid of any kinds of English test and I am afraid if the results are not good I may come home. (#1 Education student, China)

Understandably, the students did not always understand the nature of DELNA before they took the assessment, even though they are given access to information about the programme in advance. Several of them felt that the Screening did not give as good an indication of their ability as the Diagnosis, which consisted of tasks which were relatively familiar to those who had taken IELTS.

Maybe IELTS is some structure I am familiar with – I familiar with that type test but DELNA was quite different. I don't prepare enough to do right. (#14, Engineering student, Vietnam)

4.3.2 How Did They Respond to the Advisory Session and Advice?

All the participants welcomed the DELNA advisory session, where they felt "listened to" and "valued", which Chanock (2007) points out is essential. It gave them the opportunity to discuss the issues as they saw them; their strengths were acknowledged and the areas that needed improvement identified. All agreed that the advice was detailed and clear, the interview was valuable, and that they were able to keep in touch with the language adviser later.

She gave me lots of useful suggestions about my language, focus on my language. And I followed her suggestions and I received a lot of benefit. (#17, Business student, China).

The Language Adviser provided direction for English learning, suggests variable information to improve listening and writing skills. She indicated what I should study to improve – this is a good way to understand. Experts can help identify. After the discussion I know what I should improve. She gave me some documents and provided useful websites. (#18, Engineering student, Thailand)

The participants spoke about the importance of having an empathetic "expert" to talk through language issues with them, used the words "useful" and "helpful" to describe the session, and some attributed their subsequent improvement and progress to it. On the other hand, they did not always have the opportunity to follow the advice:

She suggest me a lot of thing to do but actually I don't have much time to do all the things. (#14, Engineering student, Vietnam)

4.3.3 How Did the Participants Evaluate the Language Enrichment Activities They Were Required or Recommended to Undertake?

Credit Courses

Ten of the participants, especially those with language waivers, were required to pass a suitable credit course in their first year of study. Since most of the students in our sample were in Engineering or Science, seven of them took SCIGEN 101: Communicating for a Knowledge Society. Offered by the Science Faculty, the course develops the ability to communicate specialist knowledge to a variety of audiences, through written research summaries, oral presentations and academic posters. Even though it is primarily a first-year undergraduate course, the participants found it very helpful and relevant to their future academic and professional needs.

In SCIGEN they teach us how …how to explain, how to transfer knowledge, I mean if I am writing something I am sensing – it is really helpful. …. And then a good thing teach me how to make a poster – the important thing how to do the writing. (#2, Engineering student, China)

And also give a talk in front of about five students and two teachers. Yes, because before I give the talk I had few opportunity to speak so many people, in front of so many people. (#19 Science student, China)

Two students expressed appreciation for being able to take a taught course, especially one presented "in English style", alongside native-speaking students. Thus, there were opportunities for social contacts with others in the class, although one student from China commented that he found "few chance[s] to speak with others". One source of stress for two participants was the need to obtain at least a B grade for the course, particularly since 50 % of the grade was based on a final written exam.

Apart from SCIGEN 101, two Business students took courses in Academic English, focusing on listening and reading, and writing respectively; and a more proficient Education student enrolled in a Research Methods course, which he found "challenging", but "helpful" and "important". A few students also took a course, as recommended by their supervisor, to enhance their knowledge of a subject relevant to their doctoral research.

Workshops

The participants were advised about the workshops offered by Student Learning Services (SLS). There is an initial induction day attended by all new doctoral candidates which introduces them to the whole range of workshops available to them. It is also possible for them to book an individual session with a learning adviser, particularly for assistance with writing, at both SLS and ELE. In the interviews, they made universally favourable comments about the value of the workshops and individual consultations they participated in, as well as the ELE learning resources.

> *yeah mostly the workshop in ELE and the library is very good. I try to attend most of them but I missed some of them*. (#5 Business student, Iran)
>
> *Before I submit this I attended a Student Learning Workshop – you carry your research proposal draft and they provide a native English speaker – the person who helped me edit the proposal told me my English writing was very good – because he helped others and couldn't understand as well. He proofread – … he worked with me and some information he did not understand we rewrote*. (#18 Engineering student, Thailand)

Writing was the major focus for all these activities, given the ultimate goal of producing a doctoral thesis. However, four participants were more immediately writing conference papers with their supervisors, who were doing major editing of their work, and other supervisors were insisting on early writing outputs. This gave an urgency to the students' efforts to seek assistance in improving their writing.

Conversation Groups

Although there is no assessment of speaking skills in DELNA, the advisory session gives the adviser an opportunity to make an informal assessment. In fact international doctoral candidates are well aware of their limited oral proficiency in English; only 2 of our 20 participants felt comfortable with their spoken language at the outset. Others came with the view that it would rapidly improve in an English-speaking environment:

> *I think I will make a breakthrough after I come to English speaking country, but it's not what I hoped. Maybe it will take time*. (#4 Engineering student, China)

As discussed further below, though, most found that an immersion experience was more difficult to obtain than they expected. Thus, the DELNA advisers direct doctoral candidates to the Let's Talk sessions at ELE. These sessions are scheduled several times a week and involve informal conversation in small groups with other international students, and more recently with local students or with senior Aucklanders who come in for the occasion. A facilitator is present but steps in only as the conversation closes to give some feedback.

Nine of our participants attended Let's Talk either regularly or as often as their other commitments allowed. They described it as the chance they needed to speak English in a safe environment where they were able to understand the other students

present and assess their own improvement week by week as they tried to match up to the level of those with greater fluency.

> *Before I come to NZ I had so few opportunities to speak English but I come here and joined the Let's Talk. I have more expression practise English with other people and make good communication. It is so good, really*. (#19 Science student, China)
>
> *I have been to Let's Talk – it was great – … it is like a bunch of PhD students getting there and talk about their PhD like and how they deal with supervisors – … – people from all over the world. You can hear from people talking with different accents, different cultural backgrounds. It is very interesting. I am really interesting to learning Japanese – and actually have met some Japanese friends there*. (#17, Business student, China)

There were numerous favourable comments about the impact of these sessions on the students' confidence to communicate in English. On the other hand, one participant (#11) found that a few fluent students tended to dominate the discussion, and another came to the conclusion that the value of the sessions diminished over time:

> *The problem with Let's Talk was after a while everything was the same. I mean after a few months, and I think everyone going to Let's Talk I meet a lot of people also and they just give up. I mean after that you just don't need it*. (#12, Science student, Iran)

This of course can be seen as evidence of the success of the programme if it gave the students the confidence to move on and develop their oral skills in other ways.

Listening and Reading Skills

In the area of listening skills, the DELNA advisers recommended that students access an academic podcast site developed at another New Zealand university (http://martinmcmorrow.podomatic.com). This resource draws on the well-known TED talks (www.ted.com) and has a particular focus on academic vocabulary development as well as comprehension activities. Fourteen of the participants listened regularly and all agreed that using this podcast simultaneously increased their knowledge of New Zealand and significantly improved their listening and reading skills.

As reflected in their low DELNA scores, reading was expected to be really difficult for four of the participants, and indeed they found the reading demands of their doctoral programmes somewhat overwhelming. Two participants followed the DELNA advice to attend workshops on reading strategies and access an online programme on effective reading. Others reported no particular difficulty with reading in their disciplinary area – apart from the volume of reading they were expected to complete. Three of the students found it particularly useful to work through the recommended graded readers and other materials available at ELE.

> *I love this part* [*ELE*] *– I can borrow lots of books – not academic books – there are lots of magazines and story books – not – because in our university most of the books are very , you know, just for the research, academic books, but I can go there to … borrow some books, very nice books and they also have the great website online and we can practise in English on the website- I try to do it in the lunchtime – … so I just open their website, have a look, just a few minutes, but it's good*. (#7, Engineering student, China)

4.3.4 What Other Strategies Did the Participants Adopt to Help Them Cope with the Language Demands of Their Provisional Year?

It was a common observation by the participants that they had less opportunity to communicate with others in English on campus than they might have expected. Typically, they were assigned to an office or a lab with a small number of other doctoral candidates and tended to focus on their own projects for most of the day:

> *Because I am a PhD student and we are always isolated in our research and only the time that I have during the day is just – speaking with my friend during lunch time – it is very short. Although my partner and I speak together in English…* (#14 Engineering student, Vietnam)

The situation was exacerbated when fellow PhD students were from one's own country:

> *We spend most of our time in our office – we are quiet. About 10 sentences a day – the case is getting worse with Iranian colleagues in the office. So far no gatherings of doctoral students to present work but it is going to happen*. (#9 Engineering student, Iran)
>
> *Frankly speaking, in my research group there are most – 80% Chinese. We discuss the research details in Chinese, which is very convenient for us, so we use the most common way*. (#15 Engineering student, China)

This made it all the more important for the students to follow the recommendations of the DELNA language advisers and take up the formal language enrichment activities available on campus, which they all did to varying degrees. However, five of the participants, demonstrating the "agency" that Fotovatian (2012) describes in her study of four doctoral candidates, created their own opportunities to improve their language skills. One moved out of her original accommodation close to the university because there were too many students who spoke her L1 and went into a New Zealand family homestay. She felt that the distance from the University was offset by the welcoming atmosphere and the constant need to speak English. She kept the BBC as a favourite on her laptop in her office, and took 10 min breaks every now and then, as advised by an adviser in ELE, to listen to the news in English. In the interview her communication skills were impressive and she reported that her friends had commented on her improvement. Another engineering student, who had seemed completely disoriented on arrival in Auckland, was confident and enthusiastic in the interview. He too had moved into an English-speaking homestay and had improved his spoken English by cooking the evening meal with his host.

Three others also consciously worked on their language skills by finding ways to mix with English speakers both in and outside the university. They audited lectures and attended all seminars and social occasions organised by their Faculty and by the Postgraduate Students Association. One tried to read more of the newspaper each day and they all listened to the radio, and watched films on television.

4.3.5 What Role Did the Doctoral Supervisors Play in Encouraging Language Development?

Students reported a range of reactions from satisfaction to discontent with their supervisors but the overall comments were positive. Some of the supervisors focused particularly on language skills: "they are seriously looking for my English", "he checks my work", "he reviews my content and my English". Two of the participants felt that their supervisors began to focus too closely on their English as a result of the DELNA Diagnosis.

Because of a concern about his listening skills one student kept detailed notes of the supervision meetings to make sure that he did not miss anything and another recorded them for the same reason. Three of the supervisors spoke Mandarin with their Chinese students, while another in the same situation explained that it was important for them to speak English because that was the language the student would be writing and presenting in. On the other hand, some concerns about the student-supervisor relationship did not seem particularly language-related. One student felt that she and her supervisor had a personality clash, another thought that her supervisor just "didn't like having meetings", and a third, having trouble getting time with her supervisor, described him as "very busy".

Those participants who reported having had good relationships with their supervisors also felt that the supervisory meetings were improving as their confidence in using English increased. For students with the very low DELNA scores these meetings were initially difficult:

> *At the beginning*, *I am afraid of discussing with my supervisor and co-supervisor because I got a difficult time in understanding them and also failed to express my ideas clearly. Fortunately*, *I easily face our weekly meeting although still has a few communication issues. However*, *I have started to enjoy the studying life here*. (#17Business student, China)

Since 2013, supervisors have been able to comment on their students' language development in the provisional year report. All four of the supervisors of our participants who completed their provisional year in 2013 added their reflections, with one noting that his candidate's speaking was fine and that her writing "keeps improving". Another described his student's achievements so far: a paper for which she was the main author, several presentations in faculty seminars, and involvement in the organising committee of a conference. He noted that she would be able to write her thesis "comfortably". Two comments came from supervisors whose students had had language waivers: one commented that his student "writes better than many students and will be able to write up his work"; while the other was "happy with the improvement made". These latter students were two of the five described above who had consciously taken every opportunity within and outside the university to improve their language skills during that time.

5 Discussion

The fact that newly-enrolled candidates were required to take part in DELNA was accepted by 13 of the participants, while seven voiced certain concerns. The participant who thought IELTS meant English was "over" (cf. O'Loughlin 2008) soon discovered for herself that this was not the case. The initial anxiety about the assessment experienced by at least two of the participants could have been reduced if they had been better informed about DELNA during the induction period, and this highlights the ongoing need to ensure that all candidates engage in advance with the information resources available (see Read 2008, on the presentation of DELNA; see also Yan et al. this volume). The DELNA team must also be in regular contact with the Graduate School, supervisors, Student Learning Services and other stakeholders, to ensure that the role of the assessment is clearly understood and it continues to contribute to effective academic language development among those students who need it.

Most of the participants commented that English as learnt in their home country and English in their new environment were different. For some this was so difficult on arrival that they were depressed, as Cotterall (2011) and Fotovatian (2012) have also observed. Therefore, they were grateful to simply follow advice from DELNA to take advantage of the academic enrichment activities, which all the participants generally described as helpful and relevant. The one-on-one advisory session helped dispel their concerns by offering sympathetic and targeted support in what Manathunga (2007a) describes as "this often daunting process of becoming a knowledgeable scholar" (p. 219). The main constraint for two of the participants, whose DELNA results suggested that they needed considerable language development, was that their research commitments from early on in their candidacy limited the amount of time they could spend on language enrichment. This constraint needs to be accommodated as far as possible in the DELNA advisory discussion.

As our participants reported, by taking up the language enrichment activities recommended by the adviser and sharing the responsibility for improving their academic English ability, they improved their self-confidence and this facilitated more interaction with their local peers and further language development. They thereby created a virtuous circle which "tends to increase prospects of academic success" (Sawir et al. 2012, p. 439). Those students who took courses as part of the provisional year requirements generally benefitted from them. Although one of the students with a language waiver found the credit course and the required pass a particular burden, the others who were enrolled in courses found the input from a lecturer and interaction with other students particularly useful. This interaction, which was missing in the university experience of the six doctoral researchers in Cotterall's (2011) study, was one key to confidence building, thus leading to increased use of English and an acknowledgement by the students that their skills had improved (Seloni 2012).

The Let's Talk groups at ELE were invaluable for improving confidence in listening and enabling conversation in English in an atmosphere that encouraged unembarrassed participation (Manathunga 2007a). Even though the discussions were not particularly academic in nature, this led to greater confidence in academic contexts, which was the reported outcome of a similar programme at another New Zealand university as well (Gao and Commons 2014).

As has been shown in the research on doctoral learning, the supervisory relationship tends to be stressful when the candidate is not highly proficient in English (Knight 1999; Strauss 2012). For the majority of our participants the supervisory relationship was being managed reasonably well with what appeared to be understanding of the language challenge faced by the students. The supervisors who did not seem to have time for their students and the two who spoke the students' L1 exclusively were a concern, whereas those who gave frequent writing practice and then showed that they were taking it seriously by giving feedback on the writing were doing their students a service, as Owens (2006) and Manathunga (2014) have also observed. The same applied to supervisors who engaged in joint writing tasks with their students at an early stage – something Paltridge and Starfield (2007) see as critical.

One kind of evidence for the positive engagement of supervisors has come in the form of responses to the reports sent by the DELNA Language Adviser after the advisory session for each new candidate. One supervisor, who has English as an additional language himself, recently bemoaned the fact that something similar had not been in place when he was a doctoral candidate. Another supervisor expressed full support for the DELNA and ELE programmes and was happy to work with his student to improve all aspects of her English comprehension and writing because she was:

> *one of the best postgraduate students I have ever supervised and I think her participation in the English language enrichment programme will be of great benefit to her both in her PhD and beyond*. (PhD supervisor, Science)

This emphasises the point that assisting international doctoral students to rise to the language challenge is a key step in allowing them to fulfil their academic potential.

As previously acknowledged, this was a small study of 20 self-selected participants, who may have been proactive students already positively inclined towards DELNA. In addition, the research covers only their provisional year of study and particularly their reactions to the first months at an English-medium university. Their experiences and opinions, however, mirror those reported in the studies cited above. It would be desirable to track the students through to the completion of their degree and conduct a further set of interviews in their final year of study. More comprehensive input from the supervisors would also add significantly to the evidence on the consequences of the DELNA-related academic language enrichment that the students undertook.

6 Conclusion

The provisions for academic language enrichment at Auckland are not particularly innovative. Similar courses, workshops and advisory services are provided by research universities around the world for their doctoral candidates. What is more distinctive at Auckland is, first, the administration of a post-admission English assessment to all students, regardless of their language background, to identify those at risk of language-related difficulties in their studies. Secondly, the assessment does not function simply as a placement procedure to assign students to compulsory language courses, but leads to an individual advisory session in which the student's academic language needs are carefully reviewed in the light of their DELNA performance and options for language enrichment are discussed. The resulting report, which is received by the supervisor and the Graduate School as well as by the student, includes a blend of required and recommended actions, and the student is held to account as part of the subsequent review of their provisional year of doctoral registration. The process is intended to communicate to the student that the University takes their academic language needs seriously and is willing to devote significant resources to addressing those needs.

This study has presented evidence that participants appreciated this initiative by the University and responded with effective actions to enhance their academic language skills. At least among these students who were willing to be interviewed, the indications are that the programme boosted the students' confidence in communicating in English in ways which would enhance their doctoral studies. However, academic language development needs to be an ongoing process and, in particular, the participants in this study had yet to face the language demands of producing a thesis, communicating the findings of their doctoral research, and negotiating their entry into the international community of scholars in their field. Further research is desirable to investigate the longer-term impact of the University's initiatives in language assessment and enrichment, to establish whether indeed they contribute substantially to the ability of international doctoral students to rise to the language challenge.

> *I think English study is most challenging. It is quite hard. I don't think much in the degree is as challenging – the language*, *yeah*! (# 15, Engineering student, China)

References

Aitchison, C. (2014). Same but different: A 20-year evolution of generic provision. In S. Carter & D. Laurs (Eds.), *Developing generic support for doctoral students: Practice and pedagogy*. Abingdon: Routledge.

Aitchison, C., Catterall, J., Ross, P., & Burgin, S. (2012). 'Tough love and tears': Learning doctoral writing in the sciences. *Higher Education Research & Development, 31*(4), 435–448.

Arkoudis, S., Baik, C., & Richardson, S. (2012). *English language standards in higher education: From entry to exit*. Camberwell: ACER Press.

Basturkmen, H., East, M., & Bitchener, J. (2014). Supervisors' on-script feedback comments on drafts of dissertations: Socialising students into the academic discourse community. *Teaching in Higher Education, 19*(4), 432–445.

Benzie, H. J. (2010). Graduating as a 'native speaker': International students and English language proficiency in higher education. *Higher Education Research & Development, 29*(4), 447–459.

Braine, G. (2002). Academic literacy and the nonnative speaker graduate student. *Journal of English for Academic Purposes, 1*(1), 59–68.

Carter, S., & Laurs, D. (Eds.). (2014). *Developing generic support for doctoral students: Practice and pedagogy*. Abingdon: Routledge.

Chanock, K. (2007). Valuing individual consultations as input into other modes of teaching. *Journal of Academic Language & Learning, 1*(1), A1–A9.

Cotterall, S. (2011). *Stories within stories: A narrative study of six international PhD researchers' experiences of doctoral learning in Australia*. Unpublished doctoral thesis, Macquarie University, Australia.

Dunworth, K. (2009). An investigation into post-entry English language assessment in Australian universities. *Journal of Academic Language and Learning, 3*(1), 1–13.

Dunworth, K., Drury, H., Kralik, C., Moore, T., & Mulligan, D. (2013). *Degrees of proficiency: Building a strategic approach to university students' English language assessment and development*. Sydney: Australian Government Office for Learning and Teaching. Retrieved February 24 2016, from, www.olt.gov.au/project-degrees-proficiency-building-strategic-approach-university-studentsapos-english-language-ass

Edwards, B. (2006). Map, food, equipment and compass – preparing for the doctoral journey. In C. Denholm & T. Evans (Eds.), *Doctorates downunder: Keys to successful doctoral study in Australia and New Zealand* (pp. 6–14). Camberwell: ACER Press.

Elder, C., & von Randow, J. (2008). Exploring the utility of a web-based English language screening tool. *Language Assessment Quarterly, 5*(3), 173–194.

Fotovatian, S. (2012). Three constructs of institutional identity among international doctoral students in Australia. *Teaching in Higher Education, 17*(5), 577–588.

Fox, R. (2008). Delivering one-to-one advising: Skills and competencies. In V. N. Gordon, W. R. Habley, & T. J. Grites (Eds.), *Academic advising: A comprehensive handbook* (2nd ed., pp. 342–355). San Francisco: Jossey-Bass.

Gao, X., & Commons, K. (2014). Developing international students' intercultural competencies. In S. Carter & D. Laurs (Eds.), *Developing generic support for doctoral students: Practice and pedagogy* (pp. 77–80). Abingdon: Routledge.

Grant, B. (2003). Mapping the pleasures and risks of supervision. *Discourse Studies in the Cultural Politics of Education, 24*(2), 175–190.

Knight, N. (1999). Responsibilities and limits in the supervision of NESB research students in the Social Sciences and Humanities. In Y. Ryan & O. Zuber-Skerrit (Eds.), *Supervising postgraduates from Non-English speaking backgrounds* (pp. 15–24). Buckingham: The Society for Research into Higher Education & Open University Press.

Knoch, U. (2012). At the intersection of language assessment and academic advising: Communicating results of a large-scale diagnostic academic English writing assessment to students and other stakeholders. *Papers in Language Testing and Assessment, 1*, 31–49.

Knoch, U., & Elder, C. (2013). A framework for validating post-entry language assessments. *Papers in Language Testing and Assessment, 2*(2), 1–19.

Laurs, D. (2014). One-to-one generic support. In S. Carter & D. Laurs (Eds.), *Developing generic support for doctoral students: Practice and pedagogy* (pp. 29–32). Abingdon: Routledge.

Lutz-Spalinger, G. (2010). *Final proposal, English language proficiency: Doctoral candidates*. Unpublished report, The University of Auckland.

Manathunga, C. (2007a). Supervision as mentoring: The role of power and boundary crossing. *Studies in Continuing Education, 29*(2), 207–221.

Manathunga, C. (2007b). Intercultural postgraduate supervision: Ethnographic journeys of identity and power. In D. Palfreyman & D. L. McBride (Eds.), *Learning and teaching across cultures in higher education*. Basingstoke: Palgrave Macmillan.

Manathunga, C. (2014). Reflections on working with doctoral students across and between cultures and languages. In S. Carter & D. Laurs (Eds.), *Developing generic support for doctoral students: Practice and pedagogy* (pp. 71–74). Abingdon: Routledge.

Murray, N. (2012). Ten 'Good Practice Principles' … ten key questions: Considerations in addressing the English language needs of higher education students. *Higher Education Research & Development, 31*(2), 233–246.

Nagata, Y. (1999). Once I couldn't even spell "PhD student" but now I *are* one. In Y. Ryan & O. Zuber-Skerrit (Eds.), *Supervising postgraduates from non-English speaking backgrounds* (pp. 15–24). Buckingham: The Society for Research into Higher Education & Open University Press.

O'Loughlin, K. (2008). The use of IELTS for university selection in Australia: A case study. In J. Osborne (Ed.), *IELTS research reports* (Vol. 8, pp. 145–241). Canberra: IELTS, Australia.

Owens, R. (2006). Writing as a research tool. In C. Denholm & T. Evans (Eds.), *Doctorates Downunder: Keys to successful doctoral study in Australia and New Zealand* (pp. 175–181). Camberwell: ACER Press.

Owens, R. (2007). Valuing international research candidates. In C. Denholm & T. Evans (Eds.), *Supervising doctorates downunder: Keys to effective supervision in Australia and New Zealand* (pp. 146–154). Camberwell: ACER Press.

Paltridge, B., & Starfield, S. (2007). *Thesis and dissertation writing in a second language. A handbook for supervisors*. London: Routledge.

Read, J. (2008). Identifying academic language needs through diagnostic assessment. *Journal of English for Academic Purposes, 7*(2), 180–190.

Read, J. (2015). *Assessing English proficiency for university study*. Basingstoke: Palgrave Macmillan.

Sawir, E., Marginson, S., Forbes-Mewett, H., Nyland, C., & Ramia, G. (2012). International student security and English language proficiency. *Journal of Studies in International Education, 16*(5), 434–454.

Seloni, L. (2012). Academic literary socialization of first-year doctoral students in US: A microethnographic perspective. *English for Specific Purposes, 31*(1), 47–59.

Strauss, P. (2012). 'The English is not the same': Challenges in thesis writing for second language speakers of English. *Teaching in Higher Education, 17*(3), 283–293.

The University of Auckland. (2013a). *The University of Auckland strategic plan 2013 2020*. Auckland, New Zealand.

The University of Auckland. (2013b). *Statute and guidelines for the degree of Doctor of Philosophy (PhD)*. Auckland, New Zealand.

von Randow, J. (2013). The DELNA language advisory session: How do students respond? In C. Gera (Ed.). *Proceedings of the 2012 annual international conference of the Association of Tertiary Language Advisors Aotearoa/New Zealand 8(1–14)*. Hamilton, New Zealand.

Part IV
Issues in Assessment Design

Chapter 8
Vocabulary Recognition Skill as a Screening Tool in English-as-a-Lingua-Franca University Settings

Thomas Roche, Michael Harrington, Yogesh Sinha, and Christopher Denman

Abstract Research has shown that vocabulary recognition skill is a readily measured and relatively robust predictor of second language performance in university settings in English-speaking countries. This study builds on that research by developing an understanding of the relationship between word recognition skill and Academic English performance in English-medium instruction (EMI) university programs in English-as-a-lingua-franca (ELF) contexts. The use of a Timed Yes/No (TYN) test of vocabulary recognition skill was assessed as a screening tool in two EMI university foundation programs in an Arab Gulf State: in a metropolitan state university ($N=93$) and a regional private institution ($N=71$). Pearson correlation coefficients between the TYN test and performance on university placement and final test scores ranged between 0.3 and 0.6 across the two groups and by gender within those groups. This study indicates the TYN test measures have predictive value in university ELF settings for screening purposes. The trade-off between validity, reliability, usability and the cost-effectiveness of the TYN test in academic ELF settings are discussed with consideration of test-takers' digital literacy levels.

Keywords Screening • Academic English • Placement testing • Vocabulary recognition • English as a Lingua Franca • Digital literacy

T. Roche (✉)
SCU College, Southern Cross University, Lismore, NSW, Australia
e-mail: Thomas.Roche@scu.edu.au

M. Harrington
School of Languages and Cultures, University of Queensland, Brisbane, Australia
e-mail: m.harrington@uq.edu.au

Y. Sinha
Department of English Language Teaching, Sohar University, Al Sohar, Oman
e-mail: Yogesh@soharuni.edu.om

C. Denman
Humanities Research Center, Sultan Qaboos University, Muscat, Oman
e-mail: denman@squ.edu.om

J. Read (ed.), *Post-admission Language Assessment of University Students*,
English Language Education 6, DOI 10.1007/978-3-319-39192-2_8

1 Introduction

With increasing numbers of English as second language users studying in English-medium instruction (EMI) university programs globally,[1] a growing number of studies have focused on the relationship between students' English language proficiency and academic performance. While Bayliss and Ingram's (2006) review reports on a number of studies that did not find a significant relationship between academic English proficiency and tertiary level academic performance (e.g. Cotton and Conrow 1998), a large number of studies have shown a significant positive relationship (albeit sometimes weak or moderate) between the two: in English-medium tertiary programs in English-speaking countries (Elder et al. 2007; Humphreys et al. 2012; Wang et al. 2008); and, in English-medium university programs in countries where English is considered a foreign language (Roche and Harrington 2013; Yushau and Omar 2007). These latter settings, where English plays an important economic, social or cultural (here educational) role, are increasingly referred to as English as a Lingua Franca (ELF) contexts (Jenkins 2007, 2012; Seidlhofer 2011), which are discussed more extensively by Read (Chap. 11, this volume). Regardless of the exact nature of the link between proficiency and performance it is widely recognised that English proficiency plays an important role in academic outcomes.[2] Universities in ELF settings face a fundamental challenge to ensure that incoming students have the necessary academic English skills for success in English-medium programs of study. Pre-entry foundation programs play a particularly important role in this process, especially in ELF contexts where English education in high and middle schools can be limited, and where English is not frequently used in wider society.

Placement tests are typically used in pre-entry programs to identify the level and type of English language support students need before they can undertake English-medium university study. These tests usually assess all four domains (reading, writing, listening, speaking) or subsets of these. For example, the English Placement Test (EPT) at the University of Illinois at Urbana-Champaign (UIUC) includes oral tests, as well as a reading and writing task, to provide an accurate placement (or exemption) for international students into English for Academic Purposes (EAP) writing and pronunciation classes (Kokhan 2012). Although effective, the

[1] The United Nations Educational, Scientific and Cultural Organization (UNESCO) predicts that the number of internationally mobile students will increase from 3.4 million in 2009 to approximately 7 million by 2020, with a minimum of 50 % of these students (some 3.5 million students) undertaking English language education (UNESCO 2012 in Chaney 2013). English is currently the most frequently used language of instruction in universities around the globe (Ammon 2006; Jenkins 2007; Tilak 2011).

[2] Research also points to the importance of factors such as social connections (Evans and Morrison 2011), cultural adjustment (Fiocco 1992 cited in Lee and Greene 2007) and students' understanding of and familiarity with the style of teaching (Lee and Greene 2007) as significantly contributing to students' academic success in English-medium university programs.

comprehensive nature of such tests makes them extremely time- and resource-intensive, and the feasibility of this approach for many ELF settings is questionable.

The relative speed with which vocabulary knowledge can be measured recommends it as a tool for placement decisions (Bernhardt et al. 2004). Although the focus on vocabulary alone may seem to be too narrow to adequately assess levels of language proficiency, there is evidence that this core knowledge can provide a sensitive means for discriminating between levels of learner proficiency sufficient to make reliable placement decisions (Harrington and Carey 2009; Lam 2010; Meara and Jones 1988; Wesche et al. 1996); in addition, a substantial body of research shows that vocabulary knowledge is correlated with academic English proficiency across a range of settings (see Alderson and Banerjee 2001 for a review of studies in the 1980-1990s). The emergence of the lexical approach to language learning and teaching (McCarthy 2003; Nation 1983, 1990) reflects the broader uptake of the implications of such vocabulary knowledge research in second language classrooms.

Vocabulary knowledge is a complex trait (Wesche and Paribakht 1996). The size of an L2 English user's vocabulary, i.e. the number of word families they know, is referred to as *breadth* of vocabulary knowledge in the literature (Wesche and Paribakht 1996; Read 2004). Initial research in the field (Laufer 1992; Laufer and Nation 1999) suggested that knowledge of approximately 3000 word families provides L2 English users with 95 % coverage of academic texts, a sufficient amount for unassisted comprehension. More recent research has indicated that knowledge of as many as 8000–9000 word families, accounting for 98 % coverage of English academic texts, is required for unassisted comprehension of academic English material (Schmitt et al. 2011, 2015). Adequate vocabulary breadth is a necessary but not a sufficient precondition for comprehension.[3]

For L2 users to know a written word they must access orthographic, phonological, morphological and semantic knowledge. Studies of reading comprehension in both L1 (Cortese and Balota 2013; Perfetti 2007) and L2 (Schmitt et al. 2011) indicate that readers must first be able to recognize a word before they can successfully integrate its meaning into a coherent message. Word recognition performance has been shown in empirical studies to correlate with L2 EAP sub-skills: reading (Schmitt et al. 2011; Qian 2002); writing (Harrington and Roche 2014; Roche and Harrington 2013); and listening (Al-Hazemi 2001; Stæhr 2009). In addition, Loewen and Ellis (2004) found a positive relationship between L2 vocabulary knowledge and academic performance, as measured by grade point average (GPA), in English medium-university programs, with a breadth measure of vocabulary knowledge accounting for 14.8 % of the variance in GPA.

In order to engage in skilled reading and L2 communication, (i.e. process spoken language in real time) learners need not only the appropriate breadth of vocabulary, but also the capacity to access that knowledge quickly (Segalowitz and Segalowitz 1993; Shiotsu 2001). Successful text comprehension requires lower level linguistic

[3] Other factors have also been shown to affect reading comprehension, such as background knowledge (e.g., Pulido, 2004 in Webb and Paribakht 2015).

processes (e.g. word recognition) to be efficient, that is, fast and with a high degree of automaticity, providing information to the higher level processes (Just and Carpenter 1992). For the above reasons a number of L2 testing studies (e.g. Roche and Harrington 2013) have taken response time as an index of L2 language proficiency. In this study we focus on word recognition skill, as captured in a test of written receptive word recognition (not productive or aural knowledge) measuring the breadth of vocabulary knowledge and the ability to access that knowledge without contextual cues in a timely fashion.

2 The Study

2.1 Motivation and Rationale

Vocabulary recognition screening research to date has largely been undertaken in countries where English is spoken both in and outside of the university classroom, such as in the Screening phase of the Diagnostic English Language Needs Assessment (DELNA) at the University of Auckland (Elder and von Randow 2008). Of note here is that L2 English vocabulary knowledge research involving university student participants in EMI programs in China (Chui 2006), and Oman (Roche and Harrington 2013) indicates that students in such ELF university settings have comparatively lower levels of vocabulary knowledge than their L2 peers in English-medium university programs in countries traditionally considered L1 countries. The current study addresses a gap in the literature by assessing the use of a vocabulary recognition knowledge test as a screening tool in ELF contexts with low proficiency users.

The aim of the current study is to establish the sensitivity of a Timed Yes/No (TYN) test of English recognition vocabulary skill as a screening tool in two university English-medium foundation programs in the Arab Gulf State of Oman. The TYN vocabulary knowledge test assesses both breadth and the speed of access (i.e. the time test-takers needed to access that knowledge). As a follow-up question we also examine the extent to which vocabulary recognition skill predicted overall semester performance in the two English medium Foundation programs as reflected in Final Test scores. While the primary focus is on the development of a potential screening tool, the follow-up data also provides insight into the contribution of vocabulary recognition skill to academic performance.

The current study was undertaken at two university-managed and delivered foundation programs leading to English-medium undergraduate study in Oman. Both universities assess students' English language proficiency after admission to the institution but prior to program enrolment through in-house tests to determine how much English language development, if any, those applicants are likely to need (e.g. one, two or three semesters) prior to matriculation in credit courses of study (e.g. Bachelor of Engineering). At both institutions, students who can provide evidence

of English language proficiency comparable to an overall IELTS score of 5.0 can be admitted directly into an academic credit program. Typically, over 1000 students sit comprehensive English proficiency tests prior to the start of the academic year at each institution, with students then referred to one of four options: English support in a foundation program at beginner, pre-intermediate, or intermediate level; or direct entry into an award program based on their results. Performance indicators leading to placement in those respective EAP levels are determined by test designers at each institution. Outcomes for the foundation programs are specified by the national accreditation agency (Oman Academic Accreditation Authority 2008) and are assessed through exit tests at the end of each semester. The foundation programs also cover other content, such as IT and maths skills, which are not discussed here.

As noted in the Introduction, comprehensive placement tests measuring subskills (reading, writing, listening and speaking) which provide accurate placement (or exemption) for students into foundation-level EAP programs are time- and resource- intensive. If proven effective, the TYN test assessed here may serve as an initial tool for screening students, identifying those who have the necessary English to be admitted directly to award study, and indicating for others appropriate levels of English support necessary to reach award-level study. This screening could then be followed by more comprehensive diagnostic testing as part of on-course assessment in both foundation EAP programs, and compulsory award-level English language support units. The context- independent nature of the test also provides a readily employable method for benchmarking the proficiency of students from each institution for comparison with peers studying in English at other institutions nationally and internationally.

The study aims to:

1. Identify the relationship between recognition vocabulary skill measures (size and speed) and Placement as well as Final Test scores;
2. Assess those measures as predictors alone and in combination; and,
3. Evaluate the usability of the measures for administration and scoring in this ELF setting.

2.2 Participants

Participants in this study ($N=164$) were Arabic L1 using students enrolled in the English language component of general foundation programs at two institutions. They were 17–25 years old. The programs serve as pathways to English-medium undergraduate study at University A, a metropolitan national university ($N=93$), and University B, a more recently established regional private university ($N=71$) in Oman.

The primary data collection (TYN test, Screening Test) took place at the start of a 15-week semester at the beginning of the academic year. Students' consent to take

part in the voluntary study was formally obtained in accordance with the universities' ethical guidelines.

2.3 Materials

Recognition vocabulary skill was measured using an on-line TYN screening test. Two versions of the test were used, each consisting of 62 test items. Items are words drawn from the 1,001st-4,000th most frequently occurring word families in the British National Corpus (BNC) in Part A; and from the 1 K, 2 K, 3 K, and 5 K frequency bands in Test B. Test A therefore consists of more commonly occurring words, while Test B includes lower frequency 5 K items thought to be facilitative in authentic reading (Nation 2006), and an essential part of academic study (Roche and Harrington 2013; Webb and Paribakht 2015). The composition of the TYN tests used here differed from earlier TYN testing research by the authors (Roche and Harrington 2013; Harrington and Roche 2014), which incorporated a set of items drawn from even lower frequency bands, i.e., the 10 K band. Previous research (Harrington 2006) showed that less proficient learners found lower frequency words difficult, with recognition performance near zero for some individuals. In order to make the test more accessible to the target group, the lowest frequency band (i.e. 10 k) was excluded. During the test, items are presented individually on a computer screen. The learners indicate via keyboard whether they know each test item: clicking the right arrow key for 'yes', or the left arrow key for 'no'. In order to control for guessing, the test items consist of not only 48 real word prompts (12 each from four BNC frequency levels) but also 12 pseudoword prompts presented individually (Meara and Buxton 1987). The latter are phonologically permissible strings in English (e.g. *stoffels*). The TYN test can be administered and scored quickly, and provides an immediately generated, objectively scored measure of proficiency that can be used for placement and screening purposes.

Item accuracy and response time data were collected. Accuracy (a reflection of size) was measured by the number of word items correctly identified, minus the number of pseudowords the participants claimed to know (Harrington 2006). Since this corrected for guessing score can result in negative values, 55 points were added to the total accuracy score (referred to as the *vocabulary score* in this paper). Participants were given a short 12-item practice test before doing the actual test. Speed of response (referred to as *vocabulary speed* here) for individual items was measured from the time the item appeared on the screen until the student initiated the key press. Each item remained on the screen for 3 s (3000 ms), after which it was timed out if there was no response. A failure to respond was treated as an incorrect response. Students were instructed to work as quickly and as accurately as possible. Instructions were provided in a video/audio presentation recorded by a local native speaker using Modern Standard Arabic. The test was administered using *LanguageMap*, a web-based testing tool developed at the University of Queensland, Australia.

Placement Tests. Both of the in-house placement tests included sections on reading, writing and listening, using content that was locally relevant and familiar to the students. A variety of question types, including multiple-choice, short answer, as well as true-and-false items, were used.

Final Test. Overall performance was measured by results on end of semester exams. These determine whether students need to take another semester of EAP, or if they are ready to enter the undergraduate university programs. At both institutions the Final Test mirrors the format of the Placement Tests. All test items are based on expected outcomes for the foundation program as specified by the national accreditation agency (Oman Academic Accreditation Authority 2008).

2.4 Procedure

Placement tests were administered in the orientation week prior to the start of the 15-week semester by university staff. The TYN tests were given in a computer lab. Students were informed that the test results would be added to their record along with their Placement Test results, but that the TYN was not part of the Placement Test. The test was introduced by the first author in English and then video instructions were given in Modern Standard Arabic. Three local research assistants were also present to explain the testing format and guide the students through a set of practice items.

The Final Tests were administered under exam conditions at both institutions at the end of the semester. Results for the Placement and Final Tests were provided by the Heads of the Foundation programs at the respective universities with the participants' formal consent. Only overall scores were provided.

3 Results

3.1 Preliminary Analyses

A total of 171 students were in the original sample, with seven removed for not having a complete set of scores for all the measures. Students at both institutions completed the same TYN vocabulary test but sat different Placement and Final Tests depending on which institution they attended. As such, the results are presented by institution. The TYN vocabulary tests were administered in two parts, I and II, to lessen the task demands placed on the test-takers. Each part had the same format but contained different items (see above). Reliability for the vocabulary test was measured using Cronbach's alpha, calculated separately for the two parts. Analyses were done by word (N = 48), and pseudoword (N = 14) items for the vocabulary score and vocabulary speed measures (see Harrington 2006). Each response type is

assumed to represent a different performance dimension. For vocabulary score performance on words, the reliability coefficients were: Part A = 0.75 and Part B = 0.77; for pseudowords, Part A = 0.74 and Part B = 0.73. Vocabulary speed response reliability for the words was Part A = 0.97 and Part B = 0.97, and for the pseudowords, Part A = 0.92 and Part B = 0.91. This indicates that test-taker response times were acceptably consistent on items within each test version. The reliability coefficients are satisfactory for both the vocabulary score and vocabulary speed measures.

Prior to the main analysis, performance on the respective parts was compared for a possible order effect. This was done first for the vocabulary size scores and then the vocabulary speed measures. For University A, the mean vocabulary scores (and standard deviations) were: Part I = 99.80 (18.93) and Part II = 98.87 (16.19). A paired t-test indicated that the difference in the means was not significant $t(92) = 0.338$, $p = 0.736$ (two-tailed). For University B, the statistics were: Part I = 81.20 (33.26) and Part II = 78.03 (26.97). Again the difference was not significant, $t(70) = 0.81$. $p = 0.417$. Given the lack of difference between the mean scores on the two parts, they were averaged into a single score for the respective groups.

The vocabulary speed data is reported in milliseconds by mean (and standard deviation). For University A, the figures for Part I were 1393 (259) and for Part II, 1234 (222). The 160 ms. difference between the two parts was significant, $t(92) = 8.45$ $p < 0.001$ (two-tailed), $d = 0.86$; with the latter statistic indicating a large effect for order. A similar pattern emerged for University B: Part I = 1543 (396) and Part II = 1415 (356). The mean difference was significant $t(70) = 4.18$, $p < 0.001$, $d = 0.34$. The effect size here is in the small to medium range. Response time performance was significantly faster in Part II for both groups. However, to facilitate the presentation of the placement finding, an average of the two vocabulary speed measures will be used in parallel with the vocabulary breadth measures. The vocabulary speed differences will be addressed in the Discussion.

The descriptive statistics for the test scores and vocabulary measures are presented in Table 8.1. Vocabulary measures consist of accuracy, response time and false alarms, the latter being the percentage of incorrect ('Yes') responses to pseudowords. The individual's false alarm rate is used to adjust the overall accuracy score by subtracting the proportion of 'Yes' responses to pseudowords from the overall proportion of 'Yes' responses to words (Harrington 2006). The false alarm rate alone also indicates the extent to which the test-takers were prone to guessing. An initial analysis showed a high false alarm rate for the University B students, which in turn differed by gender. As a result, the descriptive statistics are presented both by university and gender.

Table 8.1 presents the descriptive statistics for the TYN vocabulary, placement test and false alarm scores by university and gender. At both universities the male respondents typically underperformed when compared with their female peers. At University B, the regional private university, female participants had slightly better scores on all measures when compared with their male peers. For University A, the metropolitan state university, participants performed better on the vocabulary test measures (vocabulary score, vocabulary speed, false alarms) than University B participants.

Table 8.1 Descriptive statistics for TYN vocabulary, vocabulary speed, placement test and false alarm scores by university and gender

			Range		Score		Score confidence intervals (95%)	
			Low	High	*M*	*SD*	Lower	Upper
Vocabulary test (Total points = 155)	Uni A	Female n = 63	72	128	98.97	11.56	96.05	101.88
		Male n = 30	76	125	101.16	18.12	95.75	104.47
		Total N = 93			99.34	11.55	96.96	101.71
	Uni B	Female n = 37	27	122	88.75	21.52	81.57	95.92
		Male n = 34	14	119	69.68	25.99	60.61	78.75
		Total N = 71			79.62	25.47	73.59	85.64
Response time	Uni A	Female	967	1910	1335	213	1281	1388
		Male	739	1648	1270	214	1179	1360
		Total			1314	223	1268	1360
	Uni B	Female	953	2336	1390	294	1292	1488
		Male	836	2185	1575	391	1439	1712
		Total			1479	354	1395	1563
False alarms	Uni A	Female	4	68	28.69	14.24	25.10	32.27
		Male	0	64	23.69	13.96	18.48	28.90
		Total			27.07	14.27	24.13	30.01
	Uni B	Female	00	82	32.24	20.75	25.32	39.16
		Male	11	86	49.69	19.47	42.98	56.48
		Total			40.59	21.84	35.42	45.77
Placement score % (Uni specific)	Uni A	Female	41	65	54.93	5.64	53.26	56.60
		Male	39	65	54.10	7.07	51.46	56.74
		Total			54.66	6.75	53.27	56.05
	Uni B	Female	3	90	46.65	27.54	37.47	55.83
		Male	0	78	24.50	26.53	15.24	33.76
		Total			36.04	29.08	29.16	42.93
End of semester test % (Uni specific)	Uni A	Female	60	86	75.80	5.35	74.44	78.00
		Male	62	84	75.15	5.64	73.18	77.12
		Total			75.59	5.43	74.48	77.71
	Uni B	Female	35	86	73.37	10.99	70.50	76.24
		Male	21	85	61.18	12.86	58.18	64.17
		Total			67.53	13.33	64.37	70.68

Uni A University A, a national public university; *Uni B* University B, a private regional university

3.2 *False Alarm Rates*

The false alarm rates for both University A males and females (24 % and 29 %, respectively), as well as the University B females (32 %) were comparable to the mean false alarm rates of Arab university students (enrolled in first and fourth year of Bachelor degrees) reported in previous research (Roche and Harrington 2013). They were also comparable to, though higher than, the 25 % for beginners and 10 % for advanced learners evident in the results of pre-university English-language pathway students in Australia (Harrington and Carey 2009). As with other TYN studies with Arabic L1 users in academic credit courses in ELF university contexts (Roche and Harrington 2013; Harrington and Roche 2014), there were students with extremely high false alarm rates that might be treated as outliers. The mean false alarm rate of nearly 50 % by the University B males here goes well beyond this. The unusually high false alarm rate for this group, and the implications it has for the use of the TYN test for similar populations, will be taken up in the Discussion.

The extremely high false alarm level for the University B males and some individuals in the other groups means the implications for Placement and Final Test performance must be interpreted with caution. As a result, two sets of tests evaluating the vocabulary measures as predictors of test performance were performed. The tests were first run on the entire sample and then on a trimmed sample in which individuals with false alarm rates that exceeded 40 % were removed. The latter itself is a very liberal cut-off level, since other studies have removed any participants who incorrectly identified pseudowords at a rate as low as 10 % (Schmitt et al. 2011). The trimming process reduced the University A sample size by 16 %, representing 19 % fewer females and 10 % fewer males. The University B sample was reduced by over half, reflecting a reduction of 17 % in the total for the females and a very large 68 % reduction for the males. It is clear that the males in University B handled the pseudowords in a very different manner than either the University B females or both the genders in University A, despite the use of standardised Arabic-language instructions at both sites.

The University A students outperformed the University B students on the vocabulary size and vocabulary speed measures. Assumptions of equality of variance were not met, so an Independent-Samples Mann-Whitney U test was used to test the mean differences for the settings. Both vocabulary score and mean vocabulary speed scores were significantly different: $U = 4840$, $p = 0.001$, $U = 6863$, $p = 0.001$, respectively. All significance values are asymptotic, in that the sample size was assumed to be sufficient statistically to validly approximate the qualities of the larger population. For gender there was an overall difference on test score $U = 4675$, $p = 0.022$, but not on vocabulary speed, $U = 7952$ ($p = 0.315$). The gender effect for test score reflects the effect of low performance by the University B males. A comparison of the University A males and females showed no significant difference between the two, while the University B female vocabulary scores ($U = 948$, $p = 0.001$) and vocabulary speed ($U = 1155$) ($p = 0.042$) were both significantly higher than those for their male peers.

Table 8.2 Bivariate correlations between vocabulary and test measures for University A and University B, complete data set

University A n = 93 University B n = 71	University	Vocab speed	Placement test	Final test
Vocabulary size	A	–0.12	0.34**	0.10
	B	–0.59**	0.65**	0.50**
Vocabulary speed	A	–	–0.11	0.04
	B	–	–0.48**	–0.33**
Placement test	A	–	–	0.08
	B	–	–	0.53**

**$p < 0.001$ All tests two-tailed

Table 8.3 Bivariate correlations between vocabulary and test measures, data set trimmed for high false alarm values (false alarm rates >40 % removed)

University A n = 78 University B n = 38	University	Vocab speed	Placement test	Final test
Vocabulary size	A	–0.13	0.37**	0.10
	B	–0.37*	0.34*	0.24
Vocabulary speed	A	–	–0.34*	0.03
	B	–	–0.31	–0.19
Placement test	A	–	–	0.12
	B	–	–	0.27

*$p < 0.05$, **p < 0.001, All tests two-tailed

3.3 Vocabulary Measures as Predictors of Placement and Final Test Performance

The sensitivity of the vocabulary measures as predictors of placement and final test performance was evaluated first by examining the bivariate correlations among the measures and then performing a hierarchical regression to assess how the measures interacted. The latter permits the respective contributions of size and speed to be assessed individually and in combination. Table 8.2 presents the results for the entire data set. It shows the vocabulary measures were better predictors for University B. These results indicate that more accurate word recognition skill had a stronger correlation with Placement Test Scores and Final Test scores for University B than for University A.

When the data are trimmed for high false-alarm rates, the difference between the two universities largely disappears (see Table 8.3). The resulting correlation of approximately 0.35 for both universities shows a moderate relationship between vocabulary recognition skill test scores and Placement Test performance. In the trimmed data there is no relationship between TYN vocabulary recognition test scores and Final Test scores.

Regression models were run to assess how much overall variance the measures together accounted for, and the relative contribution of each measure to this amount.

Table 8.4 Hierarchical regression analyses of placement test scores with vocab score and speed as predictors for complete and trimmed data sets

Predictor	R^2	Adjusted R^2	R^2 change	B	SEB	β
Complete data set						
University A						
Vocab score	0.117	0.107	0.117****	0.200	0.058	0.334**
Vocab speed	0.121	0.102	0.005	–6.051	8.813	–0.068
University B						
Vocab score	0.419	0.411	0.419****	0.739	0.130	0.567**
Vocab speed	0.432	0.415	0.012	0.642	32.14	–0.136
Trimmed data set						
University A						
Vocab score	0.134	0.123	0.134****	0.212	0.064	0.359**
Vocab speed	0.139	0.116	0.005	–5.889	9.276	–0.068
University B						
Vocab score	0.111	0.087	0.111***	0.390	0.261	0.251
Vocab speed	0.152	0.104	0.031	–6.227	50.048	–0.219

t significant at *$p<0.05$, **$p<0.001$. *F* significant at ***$p<0.05$ and ****$p<0.001$

Table 8.4 reports on the contribution of vocabulary speed to predicting the Placement Test score criterion after the vocabulary scores were entered. Separate models were calculated for University A and University B.

As expected from the bivariate correlations, the regression analysis shows that the word measures and reaction time predicted significant variance in Placement Test scores. The University A model accounted for nearly 12 % of the Placement Test score variance while the University B one accounted for over 40 %.

Table 8.5 shows the ability of the vocabulary score (and speed) to predict Final Test performance. The vocabulary measures served as predictors of the Final Test for University B, where there was a moderate correlation (0.5) between the vocabulary scores and overall English proficiency scores at the end of the semester. There was no significant correlation with TYN Test scores and Final Test scores at University A, but it is of note that the Final Test scores for this group had a truncated range, which may reflect the higher academic entry requirements and concomitant English proficiency levels of students at the national metropolitan university as discussed below in 4.1. The results not only indicate that the TYN word recognition test is a fair predictor of performance, but also reinforce the importance of the vocabulary knowledge. The results of the study at University B are a confirmation of the significant role the vocabulary knowledge of L2 English users plays in achieving success in higher education contexts, given what is already known about the importance of other factors such as social connections (Evans and Morrison 2011), cultural adjustment (Fiocco 1992 cited in Lee and Greene 2007) and students' understanding of and familiarity with the style of teaching (Lee and Greene 2007).

The regression analyses for Final Test scores also reflect the results of the bivariate correlations: the word measures and vocabulary speed predicted significant vari-

Table 8.5 Hierarchical regression analyses of final test scores with vocab score and speed as predictors for complete and trimmed data sets

Predictor	R^2	Adjusted R^2	R^2 change	B	SEB	β
Complete data set						
University A						
Vocab score	0.117	0.107		0.200	0.058	0.334**
Vocab speed	0.121	0.102	0.005	−6.05	8.81	−0.068
University B						
Vocab score	0.251	0.240		0.248	0.068	0.567**
Vocab speed	0.252	0.230	0.001	−5.881	16.891	−0.136
Trimmed data set						
University A						
Vocab score	0.010	−0.004	0.010	0.052	0.058	0.102
Vocab speed	0.011	−0.015	0.005	3.048	8.504	0.041
University B						
Vocab score	0.059	0.033	0.111	0.096	0.058	0.203
Vocab speed	0.068	0.015	−10.754	18.374	8.81	−0.103

t significant at *$p < 0.05$, **$p < 0.001$

ance in the Final Test results. The model based on the Vocabulary test measure accounted for nearly 22 % of the Final Test variance (total adjusted $R^2 = 0.219$) while the vocabulary speed word model accounted for only 9 % (0.094).

4 Discussion

The findings are consistent with previous work that indicates vocabulary recognition skill is a stable predictor of academic English language proficiency, whether in academic performance in English-medium undergraduate university programs in ELF settings (Harrington and Roche 2014; Roche and Harrington 2013), or as a tool for placement decisions in English-speaking countries (Harrington and Carey 2009). The current study extends this research, showing the predictive power of a vocabulary recognition skill test as a screening tool for English-language university foundation programs in an ELF context. It also identifies several limitations to the approach in this context.

4.1 The TYN Test as a Placement Test for Low Proficiency Learners

Vocabulary recognition skill is a good indicator of student Placement Test performance in the target ELF context. The mid 0.3 correlations between vocabulary and Placement Tests observed for both universities in the trimmed data were at the lower

end of previous findings correlating TYN test performance with academic writing skill, where correlations ranged from 0.3 to 0.5 (Harrington and Roche 2014; Roche and Harrington 2013). Higher correlations were observed at University B when the students with very high false alarm rates were included. Although the inclusion of the high false alarm data improvespredictive validity in this instance, it also raises more fundamental questions about this group's performance on the computerised test and therefore the reliability of the format for this group. This is discussed below in 4.2. The vocabulary speed means for the present research were comparable to, if not slightly faster than, those obtained in previous studies (Harrington and Roche 2014; Roche and Harrington 2013). Mean vocabulary speed was, however, found to be a less sensitive measure of performance on Placement Tests, in contrast to previous studies where vocabulary speed was found to account for a unique amount of variance in the criterion variables (Harrington and Carey 2009; Harrington and Roche 2014; Roche and Harrington 2013).

Other TYN studies in university contexts (Harrington and Carey 2009; Roche and Harrington 2013) included low frequency band items from the BNC (i.e. 10 k band), whereas the current test versions did not. Given research highlighting the instrumental role the 8000–9000 most commonly occurring words in the BNC play in reading academic English texts (Schmitt et al. 2015, 2011), it is possible that including such items would improve the test's sensitivity in distinguishing English proficiency levels between students. This remains a question for further placement research with low proficiency English L2 students.

Results show a marked difference in performance between University A and University B on all dimensions. This may reflect differences in academic standing between the groups that are due to different admission standards at the two institutions. University A is a prestigious state institution, which only accepts students who score in the top 5 % of the annual graduating high school cohort, representing a score of approximately 95 % on their graduating certificate. In contrast, University B is a private institution, with lower entry requirements, typically attracting students who score approximately 70 % and higher on their graduating high school certificate. The differences between the two groups indicate their relative levels of English proficiency. It may also be the case that the difference between groups may be due to the digital divide between metropolitan and regional areas in Oman, with better connected students from the capital at University A more digitally experienced and therefore performing better on the online TYN test. This issue is explored further in 4.2.

4.2 The TYN Test Format

Participants in the study had much higher mean false-alarm rates and lower group means (as a measure of vocabulary size) than pre-tertiary students in English-speaking countries. A number of authors have suggested that Arabic L1 users are likely to have greater difficulties with discrete-item English vocabulary tests than

users from other language backgrounds due to differences between the Arabic and English orthographies and associated cognitive processes required to read those systems (Abu Rabia and Seigel 1995; Fender 2003; Milton 2009; Saigh and Schmitt 2012). However, as research has shown that word recognition tests do serve as effective indicators of Arabic L1 users' EAP proficiency (Al-Hazemi 2001; Harrington and Roche 2014; Roche and Harrington 2013), this is unlikely to be the reason for these higher rates. As indicated in 2.3 the test format, in particular the difference between words and pseudowords, was explained in instructions given in Modern Standard Arabic. It is possible that some students did not fully understand these instructions.

The comparatively high false-alarm rates at the regional institution may also reflect relatively low levels of digital literacy among participants outside the capital. The TYN test is an on-line instrument that requires the user to first supply biodata, navigate through a series of computer screens of instructions and examples, and then supply test responses. The latter involves using the left and right arrow keys to indicate whether the presented item is a word or a pseudoword. It was noted during the administration of the test at University B that male students (the group with the highest false-alarm rates) required additional support from research assistants to turn on their computers and log-in, as well as to start their internet browsers and enter bio-data into the test interface. As recently as 2009, when the participants in this study were studying at high school, only 26.8 % of the nation's population had internet access, in comparison to 83.7 % in Korea, 78 % in Japan, and 69 % in Singapore (World Bank 2013); and, by the time the participants in the present study reached their senior year of high school in 2011, there were only 1.8 fixed (wired) broadband subscriptions per 100 inhabitants in Oman, compared to a broadband penetration of 36.9/100 in Korea, 27.4/100 in Japan, and 25.5/100 in Singapore (Broad Band Commission 2012). Test-takers in previous TYN test studies (Harrington and Carey 2009) had predominantly come from these three highly connected countries and were likely to have brought with them higher levels of computer skills and digital literacy. It is also of note that broadband penetration is uneven across Oman, with regional areas much more poorly served by the telecommunications network than the capital, where this study's national university is located, and this may in part account for the false-alarm score difference between the two universities. Previous studies in Arab Gulf States were with students who had already completed foundation studies and undertaken between 1 and 4 years of undergraduate study; they were therefore more experienced with computers and online testing (Harrington and Roche 2014; Roche and Harrington 2013) and did not exhibit such high false-alarm scores. The current study assumed test-takers were familiar with computing and the Internet, which may have not been the case. With the spread of English language teaching and testing into regional areas of developing nations, it is necessary to be context sensitive in the application of online tests. Improved results may be obtained with an accompanying test-taking video demonstration and increased TYN item practice prior to the actual testing.

The poor performance by some of the participants may also be due to the fact that the test was low-stakes for them (Read and Chapelle 2001). The TYN test was

not officially part of the existing placement test suite at either institution and was administered after the decisive Placement Tests had been taken; high-false alarm rates may be due to some participants giving acquiescent responses (i.e. random clicks that brought the test to an un-taxing end rather than reflecting their knowledge or lack of knowledge of the items presented, see Dodorico-McDonald 2008; Nation 2007). It is therefore important that test-takers understand not only how to take the test but also see a reason for taking it. For example, we would expect fewer false alarms if the test acted as a primary gateway to further study, rather than one of many tests administered as part of a Placement Test suite, or if it was integrated into courses and contributed towards students' marks.

4.3 Gender

An unexpected finding to emerge in the study was the difference in performance due to gender. The role of gender in second language learning remains very much an open question, with support both for and against gender-based differences. Studies investigating performance on discrete item vocabulary tests such as *Lex30* (Espinosa 2010) and the *Vocabulary Levels Test* (Agustín Llach and Terrazas Gallego 2012; Mehrpour et al. 2011) found no significant difference in test performance between genders. Results reported by Jiménez Catalán (2010) showed no significant difference on the receptive tests between the male and female participants, though it was noted that girls outperformed boys on productive tests. The stark gender-based differences observed in the present study are not readily explained. Possible reasons include the low-stakes nature of the test (Nation 2007; Read and Chapelle 2001), other personality or affective variables or, as noted, comparatively lower levels of digital literacy among at least some of the male students.

5 Conclusion

Testing is resource-demanding activity involving a trade-off between the time and money available and the reliability and sensitivity of the testing instruments used. Many in-house university placement tests provide detailed placement (and potentially diagnostic) information about test-takers' English language ability across a range of sub-skills (e.g., reading, speaking, listening and writing). Such tests however, have significant resource implications for institutions, taking a great deal of time to develop, administer and mark. For ELF institutions such as the ones studied here, administering upwards of 1000 placement tests at the start of each academic year, the resource implications are considerable. The TYN test is an attractive screening tool given the limited resources needed for its administration and generation of results, thereby enabling higher-education providers in ELF contexts to use their limited resources more efficiently.

Results here show that the TYN test is a fair measure of English proficiency in tertiary ELF settings, though with some qualification. It may serve an initial screening function, identifying which EAP levels (beginner, pre-intermediate and intermediate) students are best placed in, prior to more comprehensive in-class diagnostic testing, but further research is needed to identify the TYN scores which best identify placement levels. The tests' predictive ability could be improved, potentially through adding lower-frequency test items or adding another component, such as grammar or reading items, to replace the less effective higher-frequency version I of the two vocabulary tests trialled in this study.

The use of the TYN test with lower proficiency learners in a context like the one studied here requires careful consideration. Experience with implementing the test points to the importance of comprehensible instructions and the test-taker's sense of investment in the results. The findings also underscore the context-sensitive nature of testing and highlight the need to consider test-takers' digital literacy skills when using computerised tools like the TYN test. As English continues to spread as the language of tertiary instruction in developing nations, issues of general digital literacy and internet penetration become educational issues with implications for testing and assessment.

Finally, the findings here contribute to a growing body of literature emphasising the fundamental importance of vocabulary knowledge for students studying in ELF settings. In particular, it shows the weaker a student's vocabulary knowledge, the poorer they are likely to perform on measures of their academic English proficiency and subsequently the greater difficulties they are likely to face achieving their goal of completing English preparation courses on their pathway to English-medium higher education study.

Acknowledgements This research was supported by a grant from the Omani Research Council [Grant number ORG SU HER 12 004]. The authors would like to thank the coordinators and Heads of the Foundation Programs at both institutions for their support.

References

Abu Rabia, S., & Seigel, L. S. (1995). Different orthographies, different context effects: The effects of Arabic sentence context in skilled and poor readers. *Reading Psychology, 16*, 1–19.

Agustín Llach, M. P., & Terrazas Gallego, M. (2012). Vocabulary knowledge development and gender differences in a second language. *Estudios de Lingüística Inglesa Aplicada, 12*, 45–75.

Alderson, J. C., & Banerjee, J. (2001). Language testing and assessment (part 1). State-of-the-art review. *Language Testing, 18*, 213–236.

Al-Hazemi, H. (2001). Listening to the Y/N vocabulary test and its impact on the recognition of words as real or non-real. A case study of Arab learners of English. *IRAL, 38*(2), 81–94.

Ammon, U. (2006). The language of tertiary education. In E. K. Brown (Ed.), *Encyclopaedia of languages & linguistics* (pp. 556–559). Amsterdam: Elsevier.

Bayliss, A., & Ingram D. E. (2006). *IELTS as a predictor of academic language performance*. Paper presented at the Australian International Education Conference. Retrieved from http://www.aiec.idp.com.

Bernhardt, E., Rivera, R. J., & Kamil, M. L. (2004). The practicality and efficiency of web-based placement testing for college-level language programs. *Foreign Language Annals, 37*(3), 356–366.

Broad Band Commission. (2012). *The state of broadband 2012: Achieving digital inclusion for all.* Retrieved from www.boradbandcommission.org.

Chaney, M. (2013). *Australia: Educating globally*. Retrieved from https://aei.gov.au/IEAC2/theCouncilsReport.

Chui, A. S. Y. (2006). A study of the English vocabulary knowledge of university students in Hong Kong. *Asian Journal of English Language Teaching, 16*, 1–23.

Cortese, M. J., & Balota, D. A. (2013). Visual word recognition in skilled adult readers. In M. J. Spivey, K. McRae, & M. F. Joanisse (Eds.), *The Cambridge handbook of psycholinguistics* (pp. 159–185). New York: Cambridge University Press.

Cotton, F., & Conrow, F. (1998). *An investigation of the predictive validity of IELTS amongst a sample of international students studying at the University of Tasmania* (IELTS Research Reports, Vol. 1, pp. 72–115). Canberra: IELTS Australia.

Dodorico-McDonald, J. (2008). Measuring personality constructs: The advantages and disadvantages of self-reports, informant reports and behavioural assessments. *Enquire*, *1*(1). Retrieved from www.nottingham.ac.uk/shared/McDonald.pdf.

Elder, C., & von Randow, J. (2008). Exploring the utility of a web-based English language screening tool. *Language Assessment Quarterly, 5*(3), 173–194.

Elder, C., Bright, C., & Bennett, S. (2007). The role of language proficiency in academic success: Perspectives from a New Zealand university. *Melbourne Papers in Language Testing, 12*(1), 24–28.

Espinosa, S. M. (2010). Boys' and girls' L2 word associations. In R. M. Jiménez Catalán (Ed.), *Gender perspectives on vocabulary in foreign and second languages* (pp. 139–163). Chippenham: Macmillan.

Evans, S., & Morrison, B. (2011). Meeting the challenges of English-medium higher education: The first-year experience in Hong Kong. *English for Specific Purposes, 30*(3), 198–208.

Fender, M. (2003). English word recognition and word integration skills of native Arabic- and Japanese-speaking learners of English as a second language. *Applied Psycholinguistics, 24*(3), 289–316.

Harrington, M. (2006). The lexical decision task as a measure of L2 lexical proficiency. *EUROSLA Yearbook, 6*, 147–168.

Harrington, M., & Carey, M. (2009). The on-line yes/no test as a placement tool. *System, 37*, 614–626.

Harrington, M., & Roche, T. (2014). Post–enrolment language assessment for identifying at–risk students in English-as-a-Lingua-Franca university settings. *Journal of English for Academic Purposes, 15*, 37–47.

Humphreys, P., Haugh, M., Fenton-Smith, B., Lobo, A., Michael, R., & Walkenshaw, I. (2012). Tracking international students' English proficiency over the first semester of undergraduate study. *IELTS Research Reports Online Series*, *1*. Retrieved from http://www.ielts.org.

Jenkins, J. (2007). *English as a Lingua Franca: Attitudes and identity*. Oxford: Oxford University Press.

Jenkins, J. (2012). English as a Lingua Franca from the classroom to the classroom. *ELT Journal, 66*(4), 486–494. doi:10.1093/elt/ccs040.

Jiménez Catalán, R. M. (2010). Gender tendencies in EFL across vocabulary tests. In R. M. Jiménez Catalán (Ed.), *Gender perspectives on vocabulary in foreign and second languages* (pp. 117–138). Chippenham: Palgrave Macmillan.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 98*(1), 122–149.

Kokhan, K. (2012). Investigating the possibility of using TOEFL scores for university ESL decision-making: Placement trends and effect of time lag. *Language Testing, 29*(2), 291–308.

Lam, Y. (2010). Yes/no tests for foreign language placement at the post-secondary level. *Canadian Journal of Applied Linguistics, 13*(2), 54–72.
Laufer, B. (1992). Reading in a foreign language: How does L2 lexical knowledge interact with the reader's general academic ability? *Journal of Research in Reading, 15*(2), 95–103.
Laufer, B., & Nation, I. S. P. (1999). A vocabulary size test of controlled productive ability. *Language Testing, 16*(1), 33–51.
Lee, Y., & Greene, J. (2007). The predictive validity of an ESL placement test: A mixed methods approach. *Journal of Mixed Methods Research, 1*, 366–389.
Loewen, S., & Ellis, R. (2004). The relationship between English vocabulary skill and the academic success of second language university students. *New Zealand Studies in Applied Linguistics, 10*, 1–29.
McCarthy, M. (2003). *Vocabulary*. Oxford: Oxford University Press.
Meara, P., & Buxton, B. (1987). An alternative multiple-choice vocabulary test. *Language Testing, 4*, 142–145.
Meara, P., & Jones, G. (1988). Vocabulary size as a placement indicator. In P. Grunwell (Ed.), *Applied linguistics in society* (pp. 80–87). London: Centre for Information on Language Teaching and Research.
Mehrpour, S., Razmjoo, S. A., & Kian, P. (2011). The relationship between depth and breadth of vocabulary knowledge and reading comprehension among Iranian EFL learners. *Journal of English Language Teaching and Learning, 53*, 97–127.
Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol: Multilingual Matters.
Nation, I. S. P. (1983). Learning vocabulary. *New Zealand Language Teacher, 9*(1), 10–11.
Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Boston: Heinle & Heinle.
Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review/La Revue Canadienne des Langues Vivantes, 63*(1), 59–81.
Nation, I. S. P. (2007). Fundamental issues in modelling and assessing vocabulary knowledge. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 35–43). Cambridge: Cambridge University Press.
Oman Academic Accreditation Authority. (2008). *The Oman academic standards for general foundation programs.* Retrieved from http://www.oac.gov.om/.
Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11*(4), 357–383.
Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning, 52*(3), 513–536.
Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition and testing* (pp. 209–227). Amsterdam: John Benjamins.
Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing, 18*(1), 1–32.
Roche, T., & Harrington, M. (2013). Recognition vocabulary knowledge as a predictor of academic performance in an English-as-a-foreign language setting. *Language Testing in Asia, 3*(12), 133–144.
Saigh, K., & Schmitt, N. (2012). Difficulties with vocabulary word form: The case of Arabic ESL learners. *System, 40*, 24–36. doi:10.1016/j.system.2012.01.005.
Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95*, 26–43. doi:10.1111/j.1540-4781.2011.01146.x.
Schmitt, N., Cobb, T., Horst, M., & Schmitt, D. (2015). How much vocabulary is needed to use English? Replication of Van Zeeland & Schmitt (2012), Nation (2006) and Cobb (2007). Language Teaching, 1–15.
Segalowitz, N., & Segalowitz, S. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence from second language word. *Applied Psycholinguistics, 14*, 369–385.

Seidlhofer, B. (2011). *Understanding English as a Lingua Franca*. Oxford: Oxford University Press.

Shiotsu, T. (2001). Individual differences in L2 word recognition speed: a measurement perspective. *Bulletin of the Institute of Foreign Language Education Kurume University*, *8*, 63–77. Retrieved from http://www.lognostics.co.uk

Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition, 31*, 577–607.

Tilak, J. B. G. (2011). *Trade in higher education: The role of the General Agreement on Trade in Services (GATS)* (Fundamentals of Educational Planning, Vol. 95, p. 154). Paris: UNESCO.

Wang, L., Eignor, D., & Enright, M. K. (2008). A final analysis. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 259–318). New York: Routledge.

Webb, S., & Paribakht, T. S. (2015). What is the relationship between the lexical profile of test items and performance on a standardized English proficiency test? *English for Specific Purposes, 38*, 34–43.

Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review, 53*, 13–40.

Wesche, M., Paribakht, T. S., & Ready, D. (1996). A comparative study of four ESL placement instruments. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium (LTRC), Cambridge and Arnhem* (pp. 199–209). Cambridge: Cambridge University Press.

World Bank. (2013). *Internet users (per 100 people)*. Retrieved http://data.worldbank.org/indicator/IT.NET.USER.P2.

Yushau, B., & Omar, M. H. (2007). Preparatory year program courses as predictors of first calculus course grade. *Mathematics and Computer Education, 41*(2), 92–108.

Chapter 9
Construct Refinement in Tests of Academic Literacy

Albert Weideman, Rebecca Patterson, and Anna Pot

Abstract For several reasons, the construct underlying post-entry tests of academic literacy in South Africa such as the Test of Academic Literacy Levels (TALL) and its postgraduate counterpart, the Test of Academic Literacy for Postgraduate Students (TALPS), deserves further scrutiny. First, the construct has not been further investigated in close to a decade of use. Second, acknowledging the typicality of academic discourse as a starting point for critically engaging with constructs of academic literacy may suggest design changes for such tests. This contribution surveys and critiques various attempts at identifying the typical features of academic discourse and concludes that the uniqueness of academic discourse lies in the primacy of the logical or analytical mode that guides it. Using this characteristic feature as a criterion is potentially productive in suggesting ways to add components to the current test construct of academic literacy tests that are widely used in South Africa, such as TALL, TAG (the Afrikaans counterpart of TALL), and TALPS, as well as a new test of academic literacy for Sesotho. Third, a recent analysis of the diagnostic information that can be gleaned from TALPS (Pot 2013) may inform strategies of utilising post-entry tests of language ability (PELAs) more efficiently. This contribution includes suggestions for modifications and additions to the design of current task types in tests of academic literacy. These tentative suggestions allow theoretically defensible modifications to the design of the tests, and will be useful to those responsible for developing further versions of these tests of academic literacy.

Keywords Academic literacy • Test construct • Task types • Test specifications • Academic discourse • Diagnostic information • Post-entry assessment

A. Weideman (✉) • R. Patterson • A. Pot
Office of the Dean: Humanities, University of the Free State, Bloemfontein, South Africa
e-mail: WeidemanAJ@ufs.ac.za; rebeccapatterson@gmail.com; annapot123@gmail.com

J. Read (ed.), *Post-admission Language Assessment of University Students*,
English Language Education 6, DOI 10.1007/978-3-319-39192-2_9

1 Context and Background

The effects of apartheid on education in South Africa range from the unequal allocation of resources to the continuing contestation about which language or languages should serve as medium of instruction, and at what level or levels. Inequality in education has effects, too, at the upper end of education provision, when students enter higher education.

Though increasing access to higher education since 1994 has been the norm in South Africa, the resulting accessibility has not been devoid of problems, including problems of language proficiency and preparedness of prospective new enrolments. What is more, low levels of ability in handling academic discourse are among the prime – though not the only – reasons identified as potential sources of low overall academic performance, resultant institutional risk, and potential financial wastage (Weideman 2003).

Responses to such language problems usually take the form of an institutional intervention by providers of tertiary education: either a current sub-organisational entity is tasked, or a new unit established, to provide language development courses. Conventionally, such an entity (e.g. a unit for academic literacy) would not only be established, but the arrangements for its work would be foreseen and regulated by an institutional language policy. So in such interventions two of the three prime applied linguistic artefacts, a language policy and a set of language development courses (Weideman 2014), normally come into play. How to identify which students need to be exposed to the intervention brings to the fore the third kind of applied linguistic instrument, a language assessment in the form of an adequate and acceptable test of academic language ability. These are administered either as high-stakes language tests before access is gained, or as post-entry assessments that serve to place students on appropriate language courses.

A range of post-entry tests of language ability (PELAs) in South Africa has been subjected to detailed analytical and empirical scrutiny over the last decade. These assessments include the Test of Academic Literacy Levels (TALL), its Afrikaans counterpart, the Toets van Akademiese Geletterdheidsvlakke (TAG), and a postgraduate test, the Test of Academic Literacy for Postgraduate Students (TALPS). Further information and background is provided in Rambiritch and Weideman (Chap. 10, in this volume).

2 The Continuing Importance of Construct

For post-entry tests of language ability, as for all other language assessments, responsible test design is a necessity. Responsible language test developers are required to start by examining and articulating with great care the language ability that they will be assessing (Weideman 2011). That is so because their definition of

this ability, the formulation of the hypothesized competence that will be measured, is the first critically important step to ensure that they will be measuring fairly and appropriately. What is more, it is from this point of departure – an articulation of the construct – that the technical effectiveness or validity of the design process will be steered in a direction that might make the results interpretable and useful. Without a clearly demarcated construct, the interpretation of the results of a test is impossible, and the results themselves practically useless. What is measured must inform the interpretation and meaning of the results of the measurement. These notions – interpretability and usefulness of results – are therefore two essential ingredients in what is the currently orthodox notion of test validation (Read 2010: 288; Chapelle 2012). An intelligible construct will also help to ensure that the instrument itself is relevant, appropriate, and reliable, and that its uses and impact are beneficial (Knoch and Elder 2013: 54f.).

The quest for a clear definition of the ability to be measured is complicated, however, and the ultimate choice of that definition is not devoid of compromise (Knoch and Elder 2013: 62f.). One reason for that is that some definitions of language ability may be more easily operationalisable than others. A construct has to be translated by test designers into specifications that include, amongst other things, the determination of which task types and assessment formats will be used (Davidson and Lynch 2002). It follows that test specifications must align with the definition if the test design is to be theoretically and technically defensible. Language tasks that are typical of the kind of discourse that is the target of the assessment should predominate. Yet some compromises may have to be made, not the least because test developers are constrained by any number of administrative, logistical, financial and other resource limitations, and might have to choose for their test those task types that best operationalise the construct within those constraints (Van Dyk and Weideman 2004b). The result of this may be that parts or components of a theoretically superior definition of language ability may either be overlooked or underemphasised. Messick (1998: 11) refers to this as construct under-representation, observing that "validity is compromised when the assessment is missing something relevant to the focal construct…" While a tight formulation of test specifications may curb this, some difficult design decisions might still need to be made.

A further complication that presents itself is that test designers may, once the test has been administered and used, realise that parts of it may be providing less useful or beneficial information on the ability of the candidates who sat for it, so that they require adjustment or redesign (McNamara and Roever 2006: 81f.). For example, if subtest intercorrelations are calculated (Van der Walt and Steyn 2007; Myburgh 2015: 34, 85f.), it may become evident that a pair of them (e.g. an assessment of text comprehension and interpreting a graph) may be highly correlated, and may well be testing the same component of the construct, which raises the question of whether both should be retained in subsequent designs. Or a test may have both multiple-choice questions, and an open-ended response section, that needs to be written afresh. If the results of these two sections are closely correlated, test designers may ask: how much more information is provided by the much more labour-intensive

assessment of the open-ended response section? In such cases, they are faced with a choice, therefore, of retaining the latter, perhaps for reasons offace validity, or of excluding the more labour-intensive and usually less reliable assessment, since that will cost less, without having to give up much as regards additional information about the level of language ability of candidates. In the case of the tests being referred to here, selected-response item formats were preferable for reasons of resource constraints, the necessity for a rapid turnaround in producing the results, and reliability.

The point, however, is that even the most deliberate design and careful piloting of a test is no guarantee that it will be perfect the first or even the twelfth time it is administered. As the validation processes of any test might reveal, redesign may be needed at any time, but what is being argued here is that the starting point is always the construct.

There is a third potential complication in trying to stay true to the definition of the language ability being tested, which is that new insight into the workings of language may allow us to gauge that ability better. The turn in language testing towards looking at language as communicative interaction instead of a merely formal system of knowledge (Green 2014: 173ff.; Chapelle 2012) constitutes an example of this. New perspectives on language must of necessity have an effect on what is tested, and that has certainly been the case in the tests being referred to here (Weideman 2003, 2011).

A fourth difficulty that we have encountered in test design is where the test of language ability depends on the curriculum of a national school language syllabus, as in South Africa (Department of Basic Education 2011a, b). Here the high-stakes examinations that make up the Grade 12 school exit examinations for "Home Languages" have patently over time drifted away from the original intentions of the official curricula. The intention of not only the syllabus that preceded the current syllabus for Home Languages, but also of the current version, is to emphasize communicative function. Very little of that emphasis is evident today in the three papers through which the differential, 'high-level' language ability that the curriculum calls for is being assessed (see report to Umalusi by Du Plessis et al. 2016). As this report makes clear, the best possible way to restore the integrity of these language assessments is to reinterpret the assessment with reference to the curriculum, which specifies the language construct to be tested. Without that reinterpretation, the misalignment between curriculum and the final, high-stakes assessment will endure.

Despite the possible need for compromise referred to above, or the undesirable potential of moving away from what by definition must be assessed in a test of language ability, this contribution takes as its point of departure that the clear articulation of a construct remains the best guarantee for responsible test design. This is indeed also one of the limitations of the discussion, since responsible test design patently derives from much more than a (re-) consideration of the construct. There are indeed many other points of view and design components that might well be useful points of entry for design renewal, and that will, with one exception, not be considered here, given the focus of this discussion on a single dimension: the theo-

retically defensible definition of the ability to use language in a very specific domain. With this limitation in mind, this chapter will therefore consider how a re-examination of the construct, both from the point of view of a redefinition of the target domain and from the angle of the varying emphases that such redefinition might involve, may further inform test design. It will start by looking at the development of the current construct underlying the post-entry tests which have already been identified: the Test of Academic Literacy Levels (TALL), the Toets van Akademiese Geletterdheidsvlakke (TAG, the Afrikaans version of TALL), and their postgraduate counterpart, the Test of Academic Literacy for Postgraduate Students (TALPS). Given that the construct they share has been in use for a decade, it will be argued that it deserves scrutiny at least for its being in use for some time already. Finally, this contribution will examine whether there are typical components of the ability to use academic discourse competently that might have been overlooked or underemphasised, and how that might be corrected in subsequent test designs.

3 The Current Construct and Its Theoretical Lineage

In Van Dyk and Weideman (2004a, b) there are detailed descriptions of the process of how the construct underlying the current tests referred to above was developed, and how the specifications for the blueprint of the test were arrived at. The test designers of TALL, TAG and TALPS were looking for a definition of academic literacy that was current, relevant, and reflected the use of academic discourse in a way that aligned with the notions that academics themselves have about that kind of language.

Yet finding such a construct involved a long process. The construct eventually adopted therefore derives from a developmental line that looks at language as disclosed expression (Weideman 2009), as communication, and not as an object restricted to a combination of sound, form and meaning. Moreover, as Weideman (2003) points out, it was required to take a view of the development of the ability to use academic language as the acquisition of a secondary discourse (Gee 1998). Becoming academically literate, as Blanton (1994: 230) notes, happens when

> … individuals whom we consider academically proficient speak and write with something we call *authority*; that is one characteristic — perhaps the major characteristic — of the voice of an academic reader and writer. The absence of authority is viewed as powerlessness …

How does one assess 'authority' as a measure of proficiency, however? And how does one characterise the 'academic' that stamps this authority as a specific kind? Even though the elaboration of this ability to use academic language fluently was wholly acceptable to the designers of the tests we are referring to here, the question

was how to operationalise every one of its components, formulated by Blanton (1994: 226) as the abilities of students to:

1. interpret texts in light of their own experience, and their own experience in light of texts;
2. agree or disagree with texts in light of that experience;
3. link texts to each other;
4. synthesize texts, and use their synthesis to build new assertions;
5. extrapolate from texts;
6. create their own texts, doing any or all of the above;
7. talk and write about doing any or all of the above; and
8. do numbers 6 and 7 in such a way as to meet the expectations of their audience.

Blanton's definition is noteworthy, nonetheless, because it does not define learning to become competent in academic language as knowledge merely of sound, form, and meaning. In fact, it stresses that academic discourse is communicative, interactional, and contextual. The formulations above could in principle be translated into test specifications, but were adjudged to be highly likely to result in a resource-intensive instrument. The anticipated logistical and administrative constraints prompted the test designers to look further.

A consideration of the outline of language ability in the work of Bachman and Palmer (1996) subsequently provided a second, related perspective on defining academic literacy, the ability the test designers wished to test. This approach, widely used in the field of language testing for the justification of designs, sees language ability as having two pillars: language knowledge and strategic competence (1996: 67). Whatever the advantages of this broad outline of language ability – and it also exhibits several disadvantages, not the least of which is seepage amongst the various sub-categories it proposes – it was found to be difficult, at least in the case of academic discourse, to contextualise. Specifically, it was not apparent in every case how one might fill with content the various sub-categories of these two hypothesized components of language ability.

A third perspective that could potentially solve this problem was provided by the work being done at the Alternative Admissions Research Project (AARP) at the University of Cape Town. The AARP (Yeld et al. 2000) reinterpretation of the Bachman and Palmer (1996) construct adds "understandings of typical academic tasks based largely on inputs from expert panels" (Yeld et al. 2000; cf. too Cliff and Hanslo 2005; Cliff et al. 2006). The construct is therefore enriched by the identification, amongst other things, of a number of language functions and academic literacy tasks. A streamlined version of this eventually became the final blueprint for the tests of academic literacy developed by a consortium of four multi-lingual universities, the Inter-Institutional Centre for Language Development and Assessment (ICELDA) (ICELDA 2015). This definition had as its goal the development of tests that would gauge the ability of students to

- understand a range of academic vocabulary in context;
- interpret and use metaphor and idiom, and perceive connotation, word play and ambiguity;

- understand relations between different parts of a text, be aware of the logical development of (an academic) text, via introductions to conclusions, and know how to use language that serves to make the different parts of a text hang together;
- interpret different kinds of text type (genre), and show sensitivity for the meaning that they convey, and the audience that they are aimed at;
- interpret, use and produce information presented in graphic or visual format;
- make distinctions between essential and non-essential information, fact and opinion, propositions and arguments; distinguish between cause and effect, classify, categorise and handle data that make comparisons;
- see sequence and order, do simple numerical estimations and computations that are relevant to academic information, that allow comparisons to be made, and can be applied for the purposes of an argument;
- know what counts as evidence for an argument, extrapolate from information by making inferences, and apply the information or its implications to other cases than the one at hand;
- understand the communicative function of various ways of expression in academic language (such as defining, providing examples, arguing); and
- make meaning (e.g. of an academic text) beyond the level of the sentence. (Weideman 2007: xi)

In wrestling with how academic discourse may be defined, the authors of this proposed streamlined version of the construct shared it with colleagues from a range of disciplines, in various fora and publications. The responses they received confirmed the results of the initial consultations at the time that the AARP reinterpretation was being constructed: the elements identified above not only constitute at least a number of essential components of what academic literacy entails, but resonate strongly with what academics across the disciplinary spectrum think constitutes the competent use of academic discourse (Weideman 2003).

On the basis of this refined definition of academic discourse, the designers of the tests being surveyed here began to experiment with ten different task type formats. In this their work departed from the AARP designs that emanate from essentially the same construct, since these worked with less than half that number of task types. Eventually the test designers settled on seven that fulfilled the conditions of conforming to the test construct and its detailed specifications (Van Dyk and Weideman 2004b). For the undergraduate tests, only multiple choice formats were used, for reasons already referred to above. In Table 9.1, the task types are related to the component of the construct (the basis of the specification) in the first column, with the primary task types that test these components indicated in the second column, as well as the potential, secondary task type(s) that assess them. It is clear that some of the components of the construct can potentially be measured in more than one of the possible subtests, while the reverse is also true: a subtest can potentially give insight into the ability to handle one or several components of the construct.

Table 9.1 Specifications and subtests for a test of academic literacy

Specification	Task type(s)/possible subtest(s)
Vocabulary comprehension	Knowledge of academic vocabulary
	Grammar and text relations (modified cloze)
	Understanding texts (longer reading passage)
Understanding metaphor & idiom	Understanding texts (longer reading passage)
	Text type/register
Textuality (cohesion and grammar)	Scrambled text
	Grammar and text relations (modified cloze);(perhaps) Text type / register
	Understanding texts (longer reading passage)
	Academic writing task(s)
Understanding text type (genre)	Text type/register
	Interpreting visual & graphic information
	Scrambled text
	Grammar and text relations (modified cloze)
	Understanding texts (longer reading passage)
	Academic writing task(s)
Understanding visual & graphic information	Interpreting visual & graphic information; (potentially) Understanding texts (longer reading passage)
Distinguishing essential/ non-essential	Understanding texts (longer reading passage)
	Interpreting visual & graphic information
	Academic writing task(s)
Numerical computation	Interpreting visual & graphic information
	Understanding texts (longer reading passage)
Extrapolation and application	Understanding texts (longer reading passage)
	Academic writing task(s); (possibly: Interpreting visual & graphic information)
Communicative function	Understanding texts (longer reading passage); (possibly also: Scrambled text; Grammar and text relations [modified cloze])
Making meaning beyond the sentence	Understanding texts (longer reading passage)
	Text type/register
	Scrambled text
	Interpreting visual & graphic information

4 The Typicality of Academic Discourse

While the tests that were based on this construct have now been widely scrutinised and their results subjected to empirical and critical analyses of various kinds (the 'Research' tab of ICELDA 2015 lists more than five dozen such publications), the construct referred to above has not been further investigated in close to a decade of use. In two recent studies of the construct undertaken to remedy this lack of critical engagement, Patterson and Weideman (2013a, b) take as a starting point the

typicality of academic discourse as a kind of discourse distinct from any other. They begin by tracing the idea of the variability of discourse to the sociolingual idea of a differentiated ability to use language that goes back to notions first introduced by Habermas (1970), Hymes (1971), and Halliday (1978), noting at the same time how those ideas have persisted in more current work (Biber and Conrad 2001; Hasan 2004; Hyland and Bondi 2006). Specifically, they consider how acknowledging that academic discourse is a specific, distinctly different kind of language will benefit construct renewal.

The typicality of academic discourse, viewed as a material lingual sphere (Weideman 2009: 40f.), is found to be closely aligned with the views espoused by Halliday (1978, 2002, 2003), specifically the latter's ideas of "field of discourse", genre, rhetorical mode, and register. Halliday's claim (1978: 202; cf. too Hartnett 2004: 183) that scientific language is characterised by a high degree of nominalisation, however, is deficient in several respects. First, there are other kinds of discourse (e.g. legal and administrative uses of language) in which nominalisation is also found. Second, it gives only a formal criterion for distinctness, neglecting the differences in content that can be acknowledged when one views types of discourse as materially distinct.

When one turns to a consideration of various current definitions of academic literacy, Patterson and Weideman (2013a) find that there are similar problems. In the 'critical' features of academic discourse identified by scholars such as Flower (1990), Suomela-Salmi and Dervin (2009), Gunnarsson (2009), Hyland (2011; cf. too Hyland and Bondi 2006), Livnat (2012), Bailey (2007: 10–11), and Beekman et al. (2011: 1), we find either circular definitions that identify 'academic' with reference to the academic world itself, or features that are shared across a number of discourse types. As Snow and Uccelli (2009) observe, the formally conceptualised features of academic language that they have articulated with reference to a wide range of commentators are not sufficient to define academic discourse.

Patterson and Weideman (2013a: 118) conclude that an acknowledgement of the typicality of academic discourse that is most likely to be productive is one that acknowledges both its leading analytical function and its foundational formative ('historical') dimension:

> Academic discourse, which is historically grounded, includes all lingual activities associated with academia, the output of research being perhaps the most important. The typicality of academic discourse is derived from the (unique) **distinction-making** activity which is associated with the analytical or logical mode of experience.

What is more, if one examines the various components of the construct referred to above, it is clear that the analytical is already prominent in many of them. For example, in logical concept formation, which is characterised by abstraction and analysis (Strauss 2009: 12–14), we proceed by comparing, contrasting, classifying and categorising. All of these make up our analytical ability to identify and distinguish.

5 Potential Additions to the Construct and Ways to Assess Them

As Patterson and Weideman (2013b) point out, a number of the components of the current construct already, as they should, foreground the analytical qualifying aspect of academic discourse. Which components should in that case potentially be added? Having surveyed a range of current ideas, these investigators identify the following possible additions as components of a command of academic language that can be demonstrated through the ability to:

- think critically and reason logically and systematically in terms of one's own research and that of others;
- interact (both in speech and writing) with texts: discuss, question, agree/disagree, evaluate, research and investigate problems, analyse, link texts, draw logical conclusions from texts, and then produce new texts;
- synthesize and integrate information from a multiplicity of sources with one's own knowledge in order to build new assertions, with an understanding of academic integrity and the risks of plagiarism;
- think creatively: imaginative and original solutions, methods or ideas which involve brainstorming, mind-mapping, visualisation, and association;
- understand and use a range of academic vocabulary, as well as content or discipline-specific vocabulary in context;
- use specialised or complex grammatical structures, high lexical diversity, formal prestigious expressions, and abstract/technical concepts;
- interpret and adapt one's reading/writing for an analytical/argumentative purpose and/or in light of one's own experience; and
- write in an authoritative manner, which involves the presence of an "I" addressing an imagined audience of specialists/novices or a variety of public audiences.

The first two additions may indicate the need not so much for a new task type as for a new emphasis on comparing one text with another, which is already acknowledged as a component of the construct. In some versions of TALL, for example, test takers are already expected to identify clearly different opinions in more than one text. Undoubtedly, more such comparisons are necessary to test critical insight into points of agreement and disagreement, for example. Perhaps shorter texts with contrasting opinions might also be considered, but if this ability were to be properly tested, it would add considerably to the length of a test. Otherwise, test designers might be required to ask more questions such as the following (similar to those in existing tests), which ask test takers to compare one part of a longer text with another:

> The further explanation of exactly what the author means by using the term 'development' in the first paragraph we find most clearly in paragraphs
>
> A. 2 & 3.
> B. 3 & 4.
> C. 5 & 7.
> D. 6 & 8.

or

The author discusses two divergent opinions about tapping into wind power. These opposite views are best expressed in paragraphs

A. 2 & 3.
B. 3 & 4.
C. 5 & 7.
D. 6 & 8.

We return below to the third addition: the ability to synthesize and integrate information, when we discuss another way of handling an additional writing task, based on the work of Pot (2013). A skill that, even at entry level to the academic world, relates to the notion of avoiding plagiarism and maintaining academic integrity, namely the ability to refer accurately to a multiplicity of sources, can at that lower level perhaps be measured in a task type such as the following:

References

Imagine that you have gone to the library to search for information in the form of books, articles and other material, on the topic of "Making effective presentations". You have found a number of possible sources, and have made notes from all of them for use in your assignment on this topic, but have not had the time to arrange them in proper alphabetical and chronological sequence.

Look at your notes below, then place the entry for each source in the correct order, as for a bibliography, by answering the questions below:

(a) Jay, R. 2000. How to write proposals and reports that get results . London: Pitman.
(b) Dickinson, S. Effective presentation . 1998. London: Orion Business.
(c) Hager, P.J., H.J. Scheiber & N.C. Corbin. 1997. Designing and delivering scientific, technical, and managerial presentations. New York: Wiley-Interscience.
(d) Chemical and Process Engineering, University of Newcastle-upon-Tyne. 2001. Presentation skills. Available http://lorien.ncl.ac.uk/ming/Dept/ Tips/present/present.htm.
(e) Jay, R. & A. Jay. 1994. Effective presentation: powerful ways to make your presentations more effective. Prentice-Hall: London.

The entry I placed **first** is	(a)	(b)	(c)	(d)	(e)
The entry I placed **second** is	(a)	(b)	(c)	(d)	(e)
The entry I placed **third** is	(a)	(b)	(c)	(d)	(e)
The entry I placed **fourth** is	(a)	(b)	(c)	(d)	(e)
The entry I placed **fifth** is	(a)	(b)	(c)	(d)	(e)

The entry with the **date** of publication in the wrong place is (b) (c) (d) (e)
The entry that has the **place** of publication in the wrong place is

(a) Jay (2000)
(b) Hager et al. (1997)
(c) Chemical and Process … (2001)
(d) Jay & Jay (1994)

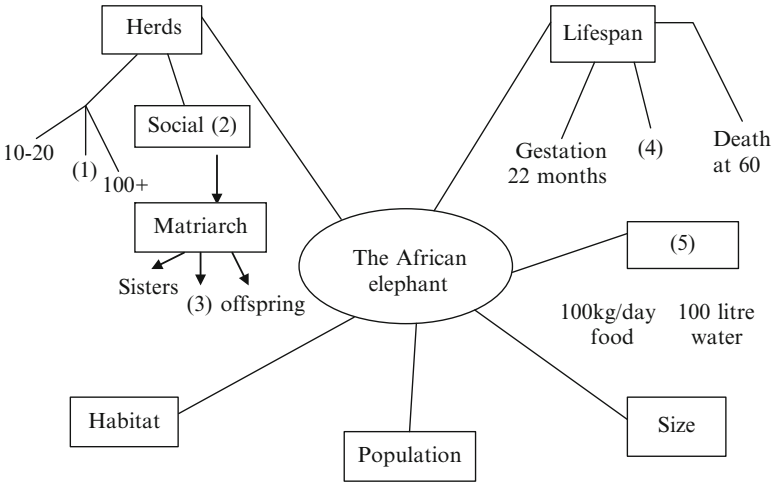


Fig. 9.1 The African elephant

The proposed fourth addition identifies an often forgotten dimension of academic work: the creativity that accompanies the visualisation of logical distinctions and concepts. Though it might again potentially add to the time taken to complete this kind of test, at least there are (as yet untested) examples of what such tasks might look like: Weideman (2006) has several suggestions to this effect that were not followed up in actual test designs. Here is one, adapted from Weideman (2007: 16–22). The text may be presented to test takers either in spoken format (as in a lecture) or in written format (Fig. 9.1):

Listen to/read the following text, look at the diagram, and then answer the questions below:

The African elephant

Elephants essentially live in herds and may be found in groups of anything between 10 and 20 or up to 50 and more, and, in rare cases, in excess of 100. Their highly developed social structure, however, remains consistent throughout. Family units are led by a cow elephant, or matriarch, and a typical family herd consists of cow elephants of various ages: the leader, and her sisters, their daughters, and their offspring.

The lifespan of an elephant is long and often eventful. For one thing, elephants …

(1) The most appropriate choice here is

(a) between 30 and 40
(b) more than 40
(c) more than 50
(d) about 70 or 80

(2) The term that is used here is

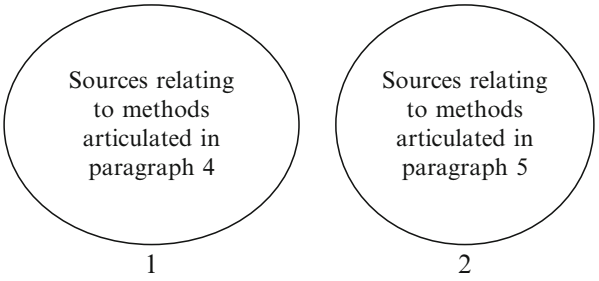
(a) development
(b) structure
(c) herd
(d) family

(3) The word that fits here is

(a) elephants
(b) age groups
(c) children
(d) daughters

Another possible question type might be to suggest a visualisation of a distinction made by the author of a text, as follows:

In the fourth and fifth paragraphs, the author makes a distinction about the different sources of various methods to generate electricity. With reference to the rest of the text, identify the circle into which the terms 'nuclear', 'coal', 'solar', 'wind', and 'water' would best fit:



1. 'nuclear' would be in circle

   A. 1.
   B. 2.
   C. neither 1 nor 2

2. 'solar' would fit into circle

   A. 1.
   B. 2.
   C. neither 1 nor 2

3. 'coal' is best categorised as belonging in circle

   A. 1.
   B. 2.
   C. neither 1 nor 2

The suggested additions under bullets five and six of the modified construct above indicate the need for a further differentiation of the tests for field or even discipline specific purposes. This is already happening in some cases: the range of tests designed by ICELDA now includes not only general tests of academic literacy, but also tests for students of disaster management, for nursing, and for financial planning. It follows that such tests must consider assessing the ability to use discipline specific terminology. This is related to the testing of the ability to use complex grammatical structures, prestigious expressions specific to a field, as well as abstract concepts and ideas. The current testing of grammar and text relations in one subtest of TALL and TALPS provides a possible format for such questions, but it should

perhaps be adapted to reflect field-specific lexical and phrasal content (Weideman and Van Dyk 2014: 95):

In the following, you have to indicate the possible *place* where a word may have been deleted, and which *word* belongs there. Here are two examples:

Charles Goodyear (1800–1860) invented the vulcanization of rubber when he was experimenting by heating a mixture of rubber and sulphur. The Goodyear story is one of either pure luck or careful research, but both are debatable. Goodyear insisted that it was [i] the [ii], though [iii] many [iv] contemporaneous [i] accounts [ii] indicate [iii] the [iv].

Where has the word been deleted?
A. At position (i).
B. At position (ii).
C. At position (iii).
D. At position (iv).

Which word has been left out here?
A. indeed
B. very
C. former
D. historically

Where has the word been deleted?
A. At position (i).
B. At position (ii).
C. At position (iii).
D. At position (iv).

Which word has been left out here?
A. historical
B. latter
C. now
D. incontrovertibly

The last two additions proposed above by Patterson and Weideman (2013b), namely the adaptation of one's reading or writing for the purposes of an academic argument, and the authoritative manner in which it should be delivered, may be less relevant for post-entry assessments at undergraduate level. At higher levels they may profitably be combined, however, so that both 'authority' and audience difference are allowed to come into play. We therefore turn next to a further consideration of how these additions might be assessed.

6 Writing with Authority

The last two additions proposed to the construct concern not only reading (finding information and evidence for the academic argument), but also writing and more specifically, writing persuasively (Hyland 2011: 177) and with authority either for a specialist or lay audience. As in the case of some of the other proposals, these additions appear to be more relevant for the discourse expected from seasoned academics than from entry-level beginners, who are normally the prime targets of post-entry language assessments. It should be noted that initially all versions of TALL did include a writing component, but the resources required to mark it reliably, as well as the high correlation between its results and that of the rest of the test resulted in

a decision to exclude it altogether. Since some wrongly equate academic literacy with the ability to write (Weideman 2013; Butler 2013), this omission would in their view constitute a potential loss of face validity for such tests (Butler 2007: 297).

The specific proposals relating to TALPS are dealt with in greater detail in another contribution (Rambiritch and Weideman, Chap. 10, in this volume). These proposals are for utilising the diagnostic information in these tests better (Pot 2013; Pot and Weideman 2015), for introducing a two-tier test consisting of a multiple-choice format first tier, and a written assignment as its second tier.

The latter refinement to the test design both for TALPS and for other tests is again motivated by a reference to the typical characteristic of academic discourse, that of making distinctions by means of language, and especially as this typical feature is embodied and expressed through argument in academic writing (Pot 2013: 58), and in the prior planning and structuring of such argument.

7 Test Refinement and Impact

From a design angle, the examination of the construct underlying post-entry tests of academic literacy in South Africa is potentially highly productive. In addition, tapping the diagnostic information the tests yield more efficiently, as well as making modifications and additions to current test task types, will provide theoretically defensible changes to their design.

The possible additions to the design of the tests referred to in this article will benefit not only the current set of assessments, but are likely to have a beneficial effect on the design of similar tests, in more languages than the current English and Afrikaans versions of the instruments. Butler and his associates at North-West University have, for example, already begun to experiment with translated versions of these post-entry assessments into Sesotho, a language widely used as first language by large numbers of students on some of their campuses, but that remains underdeveloped as an academic language (Butler 2015). A greater range of test task types will enhance the potential of the tests to provide results that are useful and interpretable, all of which may have further beneficial effects in informing policy decisions at institutions of higher learning in South Africa about an expansion of the current two languages of instruction to three, at least at some already multilingual universities.

Tests are therefore never neutral instruments, and neither is their refinement. In examining components of their design critically, and discussing modifications to them on the basis of such an examination, our goal is to continue to enhance their worth and impact. The one respect – the re-articulation of the construct – in which possible changes might be made, and that was discussed here, therefore also needs to be augmented in future research by other considerations and design principles. If Read (2010: 292) is correct in observing that the "process of test development … does not really count as a research activity in itself", we have little issue with that. Test development processes, however, are never fully complete. Once constructed,

they are subject to redesign and refinement. Since the creativity and inventiveness of test designers take precedence over the theoretical justifications of our designs, it would be a pity if the history of the rather agonising process of how to best assess a given construct were not recorded, for if that goes unrecorded, we miss the opportunity of sharing with others a potentially productive design ingredient for making or re-making tests. In more closely examining that which initially may have been secondary, namely the theoretical defence of our design, designers are, once they again scrutinise that theoretical basis, stimulated to bring their imaginations to bear on the redesign and refinement of their assessment instruments. There is a reciprocal relationship between the leading design function of an assessment measure and its foundational analytical base (Weideman 2014). This discussion has therefore aimed to provide such a record of test design, and redesign, which has been prompted by a reconsideration of the theoretical definition of what gets tested – that is, the construct.

References

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bailey, A. L. (Ed.). (2007). *The language demands of school: Putting academic English to the test*. New Haven: Yale University Press.

Beekman, L., Dube, C., & Underhill, J. (2011). *Academic literacy*. Cape Town: Juta.

Biber, D., & Conrad, S. (2001). Register variation: A corpus approach. In D. Schiffrin, D. Tannen, & H. E. Hamilton (Eds.), *The handbook of discourse analysis* (pp. 175–196). Malden: Blackwell Publishing.

Blanton, L. L. (1994). Discourse, artefacts and the Ozarks: Understanding academic literacy. In V. Zamel & R. Spack (Eds.), *Negotiating academic literacies: Teaching and learning across languages and cultures* (pp. 219–235). Mahwah: Lawrence Erlbaum Associates.

Butler, G. (2007). A framework for course design in academic writing for tertiary education. PhD thesis, University of Pretoria, Pretoria.

Butler, G. (2013). Discipline-specific versus generic academic literacy intervention for university education: An issue of impact? *Journal for Language Teaching, 47*(2), 71–88. http://dx.doi.org/10.4314/jlt.v47i2.4.

Butler, G. (2015). Translating the Test of Academic Literacy Levels (TALL) into Sesotho. To appear in *Southern African Linguistics and Applied Language Studies*.

Chapelle, C. A. (2012). Conceptions of validity. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 21–33). Abingdon: Routledge.

Cliff, A. F., & Hanslo, M. (2005). *The use of 'alternate' assessments as contributors to processes for selecting applicants to health sciences' faculties*. Paper read at the Europe Conference for Medical Education, Amsterdam.

Cliff, A. F., Yeld, N., & Hanslo, M. (2006). *Assessing the academic literacy skills of entry-level students, using the Placement Test in English for Educational Purposes (PTEEP)*. Mimeographed MS.

Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven: Yale University Press.

Department of Basic Education. (2011a). *Curriculum and assessment policy statement (CAPS) for English Home Language, Further Education and Training phase, grades 10–12*. Pretoria: Department of Basic Education.

Department of Basic Education. (2011b). *Curriculum and assessment policy statement (CAPS) for English first additional language, further education and training phase, grades 10–12*. Pretoria: Department of Basic Education.

Du Plessis, C., Steyn, S., & Weideman, A. (2016). *Towards a construct for assessing high level language ability in grade 12*. Report to the Umalusi Research Forum, 13 March 2013. Forthcoming on *LitNet*.

Flower, L. (1990). Negotiating academic discourse. In L. Flower, V. Stein, J. Ackerman, M. Kantz, K. McCormick, & W. C. Peck (Eds.), *Reading-to-write: Exploring a cognitive and social process* (pp. 221–252). Oxford: Oxford University Press.

Gee, J. P. (1998). What is literacy? In V. Zamel & R. Spack (Eds.), *Negotiating academic literacies: Teaching and learning across languages and cultures* (pp. 51–59). Mahwah: Lawrence Erlbaum Associates.

Green, A. (2014). *Exploring language assessment and testing: Language in action*. London: Routledge.

Gunnarsson, B. (2009). *Professional discourse*. London: Continuum.

Habermas, J. (1970). Toward a theory of communicative competence. In H. P. Dreitzel (Ed.), *Recent sociology 2* (pp. 41–58). London: Collier-Macmillan.

Halliday, M. A. K. (1978). *Language as social semiotic: The social interpretation of language and meaning*. London: Edward Arnold.

Halliday, M. A. K. (2002). *Linguistic studies of text and discourse*, ed. J. Webster. London: Continuum.

Halliday, M. A. K. (2003). *On language and linguistics*, ed. J. Webster. London: Continuum.

Hartnett, C. G. (2004). What should we teach about the paradoxes of English nominalization? In J. A. Foley (Ed.), *Language, education and discourse: Functional approaches* (pp. 174–190). London: Continuum.

Hasan, R. (2004). Analysing discursive variation. In L. Young & C. Harrison (Eds.), *Systemic functional linguistics and critical discourse analysis: Studies in social change* (pp. 15–52). London: Continuum.

Hyland, K. (2011). Academic discourse. In K. Hyland & B. Paltridge (Eds.), *Continuum companion to discourse analysis* (pp. 171–184). London: Continuum.

Hyland, K., & Bondi, M. (Eds.). (2006). *Academic discourse across disciplines*. Bern: Peter Lang.

Hymes, D. (1971). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 269–293). Harmondsworth: Penguin.

ICELDA (Inter-Institutional Centre for Language Development and Assessment). (2015). [Online]. Available http://icelda.sun.ac.za/. Accessed 10 May 2015.

Knoch, U., & Elder, C. (2013). A framework for validating post-entry language assessments (PELAs). *Papers in Language Testing and Assessment, 2*(2), 48–66.

Livnat, Z. (2012). *Dialogue, science and academic writing*. Amsterdam: John Benjamins.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Oxford: Blackwell.

Messick, S. (1998). *Consequences of test interpretation and use: The fusion of validity and values in psychological assessment*. Princeton: Educational Testing Service. [Online]. Available http://ets.org/Media/Research/pdf/RR-98-48.pdf. Accessed 15 Apr 2015.

Myburgh, J. (2015). *The assessment of academic literacy at pre-university level: A comparison of the utility of academic literacy tests and Grade 10 Home Language results*. MA dissertation, University of the Free State, Bloemfontein.

Patterson, R., & Weideman, A. (2013a). The typicality of academic discourse and its relevance for constructs of academic literacy. *Journal for Language Teaching, 47*(1), 107–123. http://dx.doi.org/10.4314/jlt.v47i1.5.

Patterson, R., & Weideman, A. (2013b). The refinement of a construct for tests of academic literacy. *Journal for Language Teaching, 47*(1), 125–151. http://dx.doi.org/10.4314/jlt.v47i1.6.

Pot, A. (2013). Diagnosing academic language ability: An analysis of TALPS. MA dissertation, Rijksuniversiteit Groningen, Groningen.

Pot, A., & Weideman, A. (2015). Diagnosing academic language ability: Insights from an analysis of a postgraduate test of academic literacy. *Language Matters, 46*(1), 22–43. http://dx.doi.org/10.1080/10228195.2014.986665.

Read, J. (2010). Researching language testing and assessment. In B. Paltridge & A. Phakiti (Eds.), *Continuum companion to research methods in applied linguistics* (pp. 286–300). London: Continuum.

Snow, C. E., & Uccelli, P. (2009). The challenge of academic language. In D. R. Olson & N. Torrance (Eds.), *The Cambridge handbook of literacy* (pp. 112–133). Cambridge: Cambridge University Press.

Strauss, D. F. M. (2009). *Philosophy: Discipline of the disciplines*. Grand Rapids: Paideia Press.

Suomela-Salmi, E., & Dervin, F. (Eds.). (2009). *Cross-linguistic and cross-cultural perspectives on academic discourse*. Amsterdam: John Benjamins.

Van der Walt, J. L., & Steyn, H. (2007). Pragmatic validation of a test of academic literacy at tertiary level. *Ensovoort, 11*(2), 138–153.

Van Dyk, T., & Weideman, A. (2004a). Switching constructs: On the selection of an appropriate blueprint for academic literacy assessment. *Journal for Language Teaching, 38*(1), 1–13.

Van Dyk, T., & Weideman, A. (2004b). Finding the right measure: From blueprint to specification to item type. *Journal for Language Teaching, 38*(1), 15–24.

Weideman, A. (2003). Assessing and developing academic literacy. *Per Linguam, 19*(1 and 2), 55–65.

Weideman, A. (2006). Assessing academic literacy in a task-based approach. *Language Matters, 37*(1), 81–101.

Weideman, A. (2007). *Academic literacy: Prepare to learn* (2nd ed.). Pretoria: Van Schaik Publishers.

Weideman, A. (2009). *Beyond expression: A systematic study of the foundations of linguistics*. Grand Rapids: Paideia Press.

Weideman, A. (2011). Academic literacy tests: Design, development, piloting and refinement. *SAALT Journal for Language Teaching, 45*(2), 100–113.

Weideman, A. (2013). Academic literacy interventions: What are we not yet doing, or not yet doing right? *Journal for Language Teaching, 47*(2), 11–23. http://dx.doi.org/10.4314/jlt.v47i2.1.

Weideman, A. (2014). Innovation and reciprocity in applied linguistics. *Literator, 35*(1), 1–10. http://dx.doi.org/10.4102/lit.v35i1.1074.

Weideman, A., & Van Dyk, T. (Eds.). (2014). *Academic literacy: Test your competence*. Potchefstroom: ICELDA.

Yeld, N., et al. (2000). *The construct of the academic literacy test (PTEEP)* (Mimeograph). Cape Town: Alternative Admissions Research Project, University of Cape Town.

Chapter 10
Telling the Story of a Test: The Test of Academic Literacy for Postgraduate Students (TALPS)

Avasha Rambiritch and Albert Weideman

Abstract This chapter will follow Shohamy's exhortation "to tell the story of the test" (2001). It begins by highlighting the need for the Test of Academic Literacy for Postgraduate Students (TALPS), for use primarily as a placement and diagnostic mechanism for postgraduate study, before documenting the progress made from its initial conceptualisation, design and development to its trial, results and its final implementation. Using the empirical evidence gathered, assertions will be made about the reliability and validity of the test. Documenting the design process ensures that relevant information is available and accessible both to test takers and to the public. This telling of the story of TALPS is the first step in ensuring transparency and accountability. The second is related to issues of fairness, especially the use of tests to restrict and deny access, which may occasion a negative attitude to tests. Issues of fairness dictate that test designers consider the impact of the test; employ effective ways to promote the responsible use of the test; be willing to mitigate the effects of mismeasurement; consider potential refinement of the format of the test; and ensure alignment between the test and the teaching/intervention that follows. It is in telling the story of TALPS, and in highlighting how issues of fairness have been considered seriously in its design and use that we hope to answer a key question that all test designers need to ask: Have we, as test designers, succeeded in designing a socially acceptable, fair and responsible test?

Keywords Academic literacy • Academic writing • Language tests • Postgraduate study • Intervention • Test design • Fairness • Transparency • Accountability

A. Rambiritch (✉)
Unit for Academic Literacy, University of Pretoria, Pretoria, South Africa
e-mail: Avasha.Rambiritch@up.ac.za

A. Weideman
Office of the Dean: Humanities, University of the Free State, Bloemfontein, South Africa
e-mail: WeidemanAJ@ufs.ac.za

J. Read (ed.), *Post-admission Language Assessment of University Students*,
English Language Education 6, DOI 10.1007/978-3-319-39192-2_10

1 Language and Learning in South African Tertiary Institutions

Twenty years after the demise of apartheid, South Africa is still reeling from its effects, as was indicated in the previous chapter (Weideman et al. 2016). The trauma of Bantu education, a separatist system of education, continues to reverberate through the country. Unfair and unequal distribution of resources, poor and/or unqualified teachers, and overcrowded classrooms have had long-term effects on the education system, and on the actual preparedness of (historically disadvantaged) students for tertiary study, even today. Despite the progress made in many areas, the policy of racial segregation has left fractures that will take years still to heal. One area most in need of healing remains the historically contentious issue of language and its use as a medium of instruction, especially in institutions of higher education.

Since 1994, tertiary institutions have had to deal with the challenge of accepting students whose language proficiency may be at levels that would place them at risk, leading to low pass rates and poor performance. This is a problem not specific only to students from previously disadvantaged backgrounds. Language proficiency is low even amongst students whose first languages are English or Afrikaans, which are still the main languages of teaching and learning at tertiary level in South Africa. Low levels of proficiency in English generally mean that students are not equipped to deal with the kind of language they encounter at tertiary level. For many students who have been taught in their mother tongue, entering university is their first experience of being taught in English.

Tertiary institutions, including those considered previously advantaged, today need contingency measures to deal with this situation. Not accepting students because of poor language proficiency would simply have meant a repetition of the past, since the issue of being denied access is one that is rooted in the history of the country. The trend has been to set up specific programmes to assist these students. Different institutions have, however, taken different routes. Some have set up academic support programmes, departments and units, while others have offered degrees and diplomas on an extended programme system, where the programme of study is extended by a year to ensure that the relevant academic support is provided.

At the University of Pretoria, as at other universities in the country, poor pass rates and low student success are issues of concern. The university also attracts students from other parts of Africa and the world, lending even more diversity to an already diverse environment. Mdepa and Tshiwula (2012: 27) note that in 2008 more than 9500 students from African countries outside of the 15-state Southern African Development Community (SADC) studied in South Africa. Very often these students do not have English as a first language, requiring that measures have to be put in place to assist them to succeed academically. At the University of Pretoria one such measure was put in place at the first two levels of postgraduate study (honours and master's level), by offering as intervention a module that focused

on helping students to develop their academic writing. Over time, there has been an increasing demand for the course, as supervisors of postgraduate students have recognised the inadequate academic literacy levels of their students.

Butler's (2007) study focused on the design of this intervention, a course that provides academic writing support for postgraduate students. He found that the students on the course were mostly unfamiliar with academic writing conventions, were often unable to express themselves clearly in English, and had not "yet fully acquired the academic discourse needed in order to cope independently with the literacy demands of postgraduate study" (Butler 2007: 10). Information elicited from a questionnaire administered to students and supervisors and from personal interviews with supervisors as part of Butler's study confirmed that these students experienced serious academic literacy problems, and that as a result might not complete their studies in the required time. Worrying also is the fact that the survey showed that some students (20 %) had never received any formal education in English and that a large group of them (30 % for their first degree and 44 % for honours) did not use English as a medium of instruction for their previous degrees (Butler 2009: 13–14). What became clear from the results of the study was the need for a "reliable literacy assessment instrument" that would "provide one with accurate information on students' academic literacy levels" (Butler 2007: 181).

This contribution is focused specifically on telling the story of the development and use of that test, i.e. on the design and development of what is now called the Test of Academic Literacy for Postgraduate Students (TALPS). In so doing, it takes the first step towards ensuring transparency and accountability in the design and development of such tests. We turn below first to an exposition of these two ideas, and subsequently to how they relate to interpreting and using test results, as well as the interventions – in the form of language development courses – that the use of their results implies. Finally, we consider what other tests that aim to earn a reputation of being responsibly designed might be able to learn from the intentions and the objectives of the designers of TALPS.

2 Transparency and Accountability in Language Testing

Unfair tests, unfair testing methods and the use of tests to restrict or deny access contribute to a negative attitude to tests. The move in the recent past (Shohamy 2001, 2008; Fulcher and Davidson 2007; McNamara and Roever 2006; Weideman 2006, 2009) has therefore been to promote the design and development of fair tests, by test developers who are willing to be accountable for their designs. This can be seen as part of a broader social trend whereby the concept of transparency has become the watchword in government and politics, in the corporate world, in the media and even in the humanities and social sciences. Naurin (2007) states that transparency literally means that it is possible to look into something, to see what is going on:

> A transparent organisation, political system, juridical process or market is one where it is possible for people outside to acquire the information they need to form opinions about actions and processes within these institutions. (Naurin 2007: 2)

In practical terms, transparency means that information is easily available to those who need it and that, importantly, this availability of information allows an open dialogue between those within and those outside of the organisation. The concept of transparency in the field of language testing, however, has not been comprehensively explored. Moreover, while testing experts have stressed the need for an open dialogue between test developers and test takers, for test takers to be able to ask questions about the tests and for test developers to take responsibility for their designs, this has not always happened in practice. A definition of the transparency of a test as the "availability of information about its content and workings" (Weideman 2006: 82) is a first step towards setting this right.

The term 'accountability', like the term 'transparency', features prominently in the literature of many disciplines: commerce, law, education, public management and human resources (see Norton 1997; Beu and Buckley 2004), to name just a few. According to Sinclair (1995: 220), accountability entails a relationship in which people are required to explain and take responsibility for their actions. Bovens (2005: 7) defines accountability as a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct to the forum, which then becomes a basis for further interrogation of the conduct. From this basis affected parties can pose questions and pass judgment, and even sanction the actor. The same kind of relationship is echoed in other discussions. According to Frink and Klimoski (2004: 2), for example, definitions of accountability tend to

> revolve around two specific themes. One theme concerns the context, that is, who and what is involved in a given situation, and the second theme involves the notion of an evaluation and feedback activity in some form. (Frink and Klimoski 2004: 3)

Explained simply: "Accountability involves an actor or agent in a social context who potentially is subject to observation and evaluation by some audience(s), including one's self" (2004: 3). There are also

> standards, or expectations against which the agent's behaviour are compared, and the belief on the part of the agent of some likelihood that he or she may need to answer for, justify, or defend the decisions or behaviours. In addition, it is important that there are outcomes for the agent (i.e., sanctions, rewards, or punishments that can be explicit or implicit, and also objective or subjective). (Frink and Klimoski 2004: 4)

In explaining his use of the term accountability in the field of language testing, Weideman (2006: 72) turns to the definition provided by Schuurman (2005), which stresses the need for actors to be aware of their actions and to "give account of the same to the public" (Schuurman 2005: 42).

It is clear that test developers should therefore be concerned with making information about their tests available to those most affected, and be willing to take responsibility for their designs. These issues become relevant when one works

within a framework such as that proposed by Weideman (2009), which calls for a responsible agenda for applied linguistics, to ensure that the notions of responsibility, integrity, accessibility and fairness can be articulated in a theoretically coherent and systematic way. The framework he refers to is based on a "representation of the relationship among a select number of fundamental concepts in language testing" (Weideman 2009: 241; for a more detailed exposition, cf. Weideman 2014). Weideman points out that the technical unity of multiple sources of evidence, relating for example to the reliability of a test, its validity and its rational justification, and brought together systematically in a validation argument, utilises several such foundational or constitutive applied linguistic concepts (2009: 247). These may also be designated necessary requirements for tests, and in what follows these requirements will again be highlighted with reference to the process of the design of TALPS.

In the framework of requirements employed here, the design of a test also links with ideas of its public defensibility or accountability, and the fairness or care for those taking a test (Weideman 2009: 247). In employing a set of design conditions that incorporates reference to the empirical properties and analyses of a test, as well as a concern for the social dimensions of language testing, one is able to ensure that transparency and accountability can be taken into consideration in the testing process. This contribution takes the telling of the story of the design and development of TALPS as a starting point for ensuring transparency and accountability.

2.1 Deciding on a Construct

A first step for the developers of TALPS was to find an appropriate construct on which to base the test and the intervention. Bachman and Palmer (1996: 21) define a construct as the "specific definition of an ability that provides the basis for a given test or test task and for interpreting scores derived from this task." The developers chose to base TALPS on the same construct as the Test of Academic Literacy Levels (TALL) (see Patterson and Weideman 2013; Van Dyk and Weideman 2004). The TALL, an undergraduate level test, was in that sense a sounding board for TALPS – the success of TALL was in fact one of the most important motivations for the development of the postgraduate assessment. Both these tests are designed to test the academic literacy of students, the difference being that one is directed at first year students while the other is intended for postgraduate students. For the blueprint of the test, we refer to Weideman et al. (2016 – this volume). It should be noted, however, that while the components of the construct were considered to be adequate, the developers took into account that in its operationalisation – the specification and development of task types to measure the construct – care had to be taken with both the level (postgraduate) and format (including the consideration of more and other subtests than in an undergraduate test).

2.2 Specification

The next step for the developers of TALPS was to align the construct of the test with specifications. Davidson and Lynch (2002: 4; cf. too Davies et al. 1999: 207) state that "the chief tool of language test development is a test specification, which is a generative blueprint from which test items or tasks can be produced". They observe that a well-written test specification can generate many equivalent test tasks. The discussion of specifications at this point is focused specifically on item type specification and how they align with the construct of academic literacy used for this test.

In addition to the task types (subtests) employed in TALL, it was decided to include in TALPS a section on argumentative writing. At postgraduate level it is essential that students follow specific academic writing conventions and it is important to test whether students are equipped with this knowledge. Butler (2009: 294) states: "In the development of TALPS we have also considered the importance of testing students' productive writing ability specifically (in the production of an authentic academic text), as well as their editing ability". Important for Butler (2009: 11) was the fact that if the test did not contain a section on writing, it would affect face validity. In addition to the question on writing there is a question that tests students' editing skills. Table 10.1 outlines the eight sections that now appear in TALPS, as well as a brief summary of what aspect of academic literacy each tests:

With the exception of Sect. 8, all other subtests are in multiple-choice format, as for TALL, and for the same reasons: ease of marking, reliable results, early availability of results, economical use of resources, and the availability of imaginative and creative design capacity (Van Dyk and Weideman 2004: 16). Important, as well, as pointed out by Du Plessis (2012), and which ties in closely with the need for transparency and accountability, is that test takers have the opportunity to consult sample tests before attempting to write TALPS, so gaining an understanding of how multiple-choice test items work in language tests at this level.

2.3 The Process of Development of TALPS

The process that was followed in developing TALPS saw the refinement of three drafts, with the third draft version of the test becoming the first post-refinement pilot. This (third draft) version of the test was made up of the following task types, items and marks, as indicated in Table 10.2.

Between draft one and draft three most changes were made in the Understanding texts section. In draft one this section had 45 items, in draft two it had 28 items and in the third and final draft version of the test it had 21 items, some weighted more heavily in order to achieve better alignment with what were considered to be critically important components of the construct. The four items in this section that were weighted more measured aspects relating to the student's ability to understand and

Table 10.1 The TALPS blueprint

Test section	Aspect of literacy measured
Section 1: Scrambled text	Textuality (knowledge of cohesion, grammar)
A number of sentences that need to be re-organized into a coherent passage	Understanding and responding to the communicative function of the text
Section 2: Interpreting graphs and visual information	Understanding genres
A short text passage and accompanying graph requiring numerical calculations and visual inferences	Visual literacy
	Interpreting of information
	Extrapolation and application of information
Section 3: Academic vocabulary	Advanced vocabulary knowledge Understanding and responding to the communicative function of the text
This section includes vocabulary items based on Coxhead's (2000)Academic Word List, mainly from the selection of less frequently used words	
Section 4: Text types	Understanding genres
A selection of phrases and sentences representing different genres which have to be matched with a second group of phrases and sentences	Identifying registers
	Making meaning beyond sentence level
Section 5: Understanding texts	Critical thinking
A lengthy reading passage and series of questions to be answered	Understanding and responding to the communicative function of the text
	Deriving meaning beyond sentence level Extrapolating and applying information Distinguishing essential/non-essential information
	Drawing conclusions and making inferences
Section 6: Grammar and text relations	Meaning making
A variation of cloze procedure in which certain words are deleted from a text	Understanding and responding to the communicative function of the text Knowledge of cohesion
Section 7: Text editing	Knowledge of syntax
A passage in which a number of grammatical errors have been made requiring correction	Knowledge of morphology
	Knowledge of semantics
Section 8: Academic writing	Ability to synthesize texts
A short z on information provided in the test	Making meaning beyond the level of the sentence
	Interpreting information
	Understanding and responding to the communicative function of the text
	Extrapolation and application of facts
	Knowledge of genres and registers
	Applying coherence
	Referencing

(Du Plessis, 2012: 53)

Table 10.2 Table of task types, items and marks in Draft 3

Task type	Items	Marks
Scrambled text	5	5
Graphic and visual literacy	10	10
Dictionary definitions	–	–
Academic vocabulary	10	10
Text type	5	5
Understanding texts	**21**	**25**
Grammar and text relations	15	15
Text editing	10	10
Total	**76**	**80**

(Geldenhuys, 2007: 78)

respond to communicative functions in the text, to extrapolate and to apply information, as well as the ability to draw conclusions and make inferences. The reason for the additional marks for communicative function, extrapolation, and inferencing is that it is not always possible to find a text (for comprehension type questions) that is fully exploitable, and in just the right measure, for all components of the construct. Hence the marks relating to questions that assess such components should be weighted differently than, say, questions that are abundantly represented, for example, those on relations between different parts of the text (assessing insight into cohesion and coherence).

Though piloted in previous versions, the section on Dictionary definitions was left out of this final version of the test. According to the descriptive statistics of the drafts of TALPS, the Dictionary definitions subtest had a mean p-value of 84.2 in the pilots, meaning that a high percentage of the test population got these correct. Items that are too easy or too difficult are less likely to contribute to the overall ability of a test to discriminate (Davies et al. 1999:95), which justifies their removal here. It also deserves to be noted that the first pilot of the test was carried out on first-year students at the University of Pretoria who were registered for the compulsory academic literacy intervention. For ease of administration, it did not yet include the subtest requiring students to write an argumentative text.

The subsequent pilot of this final draft version of the test was carried out in September 2007 on two groups of postgraduate students at North-West University (NWU) and the University of Pretoria (UP). This version of the test included the section on academic writing, with the question to be answered requiring an argument to be constructed, with appropriate acknowledgment and referencing, from information in the texts used in the other subtests.

An analysis of the results of the TALPS pilots yields valuable information relating to the quality and efficiency of the test. Importantly, the primary purpose of such piloting and ongoing refinement is to construct an initial picture of test validity and reliability (Second Language Testing Inc 2013). Despite being essential parts of the validation narrative, reliability and validity do not tell the entire story of the test, and while the focus of this study is to identify the other equally important parts of the tale, issues of reliability and validity are where the story begins.

The statistics package used (TiaPlus: cf. CITO 2006) provides us with two measures of test consistency: Cronbach's alpha and Greatest Lower Bound (GLB). All pilots of the test rendered very impressive reliability measures. The first pilot had a reliability of 0.85 (Cronbach's alpha) and 0.92 (GLB). One pre-final draft had measures of 0.93 (Cronbach's alpha) and 1.00 (GLB). The final version of the test had measures of 0.92 (Cronbach's alpha) and 0.99 (GLB). In the TALPS final version, the standard error of measurement for the combined group of students is at 3.84.

One other statistical measure rendered by the package used is the average *Rit*-values or the discriminative ability of the test items. One of the main purposes of a test is to be able to discriminate between the test-takers (Kurpius and Stafford 2006: 115). The mean *Rit*-values for the third pilot are relatively stable at 0.40, which is well above the 0.30 benchmark chosen. In addition, the variance around the mean seems to be quite stable, suggesting a normal or even distribution of scores around the mean.

The need for validation at the a priori stage of test development (Weir 2005) is provided for here in this initial evidence that TALPS is a highly reliable test What is more, the test is based on a theoretically defensible construct, and an argument can be made for the close alignment of that construct and the test items. The *Rit*-values of the items further indicate that the test discriminates well between test-takers. Finally, the internal correlations of the different test sections satisfy specific criteria and the face validity of the test meets the expectations of potential users. As regards the latter, the observations supporting Butler's (2009) study persuaded the test designers that it would be difficult to promote a test, especially among postgraduate supervisors, a prime group of users of the results of the test, if it did not have a subsection assessing academic writing. While acknowledging that this open-ended format would be less easy to administer than a selected-response format, the test developers believed that the evidence indicated that the inclusion of a writing subtest in the test would without doubt enhance its intuitive technical appeal or face validity (for a further discussion, see Du Plessis 2012: 65). These initial observations – on reliability, alignment with the construct, ability to discriminate, subtest intercorrelations, and face validity - provide the first pieces of evidence for a systematic and integrated argument in the process of validating TALPS.

However, this is just the first part of the tale. A fair and responsible test is one that is not only valid and reliable, and for which validation evidence can be systematically presented, but socially acceptable as well. McNamara and Roever's (2006: 2) observation that a "psychometrically good test is not necessarily a socially good test" is relevant here, because a core concern in test design is the social responsibility that the test developers have, not just to the test takers (postgraduate students) but to everyone affected by the test – supervisors, parents, test administrators and society at large. Experts acknowledge that language testing cannot take place in isolation, without reference to context, or excluding the very people whose lives are most affected by the use of test scores. Shohamy (1997, 2001), McNamara and Roever (2006), Bachman and Palmer (1996), and Hamp-Lyons (2000a, b, 2001), among others, have pointed out the negative ways in which tests have been used, and the negative effects these have had on test takers. In addition, they have discussed what

they believe test developers (and test takers) should do to ensure the development and use of fair and responsible tests. In the hope of contributing to our understanding of responsible test design, the next part of this narrative deals with issues related to the social dimensions of language testing.

3 Towards Transparency and Accountability in the Design and Use of TALPS

The transparency of a test, as explained earlier, refers to the availability of information about its content and workings (Weideman 2006: 82). True transparency, however, means that this information should be understandable, not just to experts in the field, but to a lay audience as well. The narrative here, documenting the story of the design and development of TALPS, as well as the conference papers and academic articles being referred to, provides information that is easily accessible to other academics interested in the design and use of the assessment. The real concern, however, is to ensure that the information is accessible and understandable to other audiences and affected parties who are interested in the use of the test. With regards to TALPS a first important step was to ensure that information about the test was available to prospective test takers in the form of brochures and on the websites of the universities using the test. These information media anticipate and attempt to answer all manner of questions test takers may have about the test.

In addition to this, prospective students are referred to the ICELDA (Inter-institutional Centre for Language Development and Assessment) website (ICELDA 2015), which allows them access to a number of sample tests. While the tests cannot be printed or downloaded, students can spend time acquainting themselves with the format of the test and, if interested, can complete the test or any number of items they might wish to attempt. Providing students with a sample of the test is one way of ensuring transparency. Very often what is most daunting about taking a test is the fact that the test taker does not know what to expect. While providing examples of tests is now almost routine internationally, the range of samples of academic literacy tests provided by ICELDA is the only example of such practice in South Africa. Other high-stakes tests of academic literacy in South Africa which are often used for undergraduate access to higher education (such as the academic and quantitative literacy test of the National Benchmark Tests) do not make such samples available, and in fact this has made the examples offered by ICELDA by far the most popular subpage to visit on their website. It was exactly the pressure on that organisation to provide downloadable or printable examples (instead of the secured online versions) that has led to the publication of two inexpensive books of sample tests, one in English containing six sample tests at various levels (Weideman and Van Dyk 2014), and another in Afrikaans.

These measures go some way in making as much information as possible available about the test. Importantly, the test designers become accessible, not distant

experts who may be tempted to hide behind the "scientific" authority of their designs (Weideman 2006: 80). Full contact details of the designers are available on the ICELDA website, allowing any prospective user or test taker the opportunity to make contact should they choose to. In Du Plessis's (2012) study, there is a fairly comprehensive survey of students' perceptions of TALPS, and she finds that, although "much can be done to increase transparency about the nature and purpose of the postgraduate literacy test, the results of the survey do not support the initial hypotheses that students would be predominantly negative towards the TALPS and that its face validity would be low" (Du Plessis 2012: 122). The test takers overwhelmingly agreed (directly after they had taken the test, but before their results were known) that they thought the test would measure accurately and fairly. In addition to this, and in line with the further suggestions made in this study for increasing transparency, the test has been promoted effectively within the institutions where it is used. Presentations have been made to faculties about the value of the test and the intervention programme. As was the case with TALL, the test developers have published articles about TALPS in scholarly journals (see Rambiritch 2013), in addition to presenting papers/seminars at national and international conferences. By doing this, the test developers have sought the opinions of other experts in the field. As has been pointed out above, the opinions of prime users of the test, postgraduate supervisors (cf. Butler 2009), had already been sought at the design stage, especially when considering the face validity of the test. A test of this nature will constantly need refinement, and such refinement is stimulated by its being evaluated by others working in the same field.

The level of transparency and openness about TALPS – whether it is as yet adequate or not - that has been achieved is a prerequisite for satisfying concerns related to accountability and the need to take responsibility for its design. In the case of applied linguists and test designers the challenge, however, is always to be "doubly accountable" (Bygate 2004: 19), that is, accountable to peers within the discipline within which they work, and accountable to the communities they serve.

This need to publicly defend the design, or public accountability in language testing, has been referred to by many in the field. Boyd and Davies (2002) call for the profession of language testing to have high standards, with members who are conscious of their responsibilities and open to the public (2002: 312), noting that it is not too late for language testers to "build in openness to its professional life" (2002: 312). Rea-Dickins (1997: 304) sees a need for healthy and open relationships among all stakeholders (learners, teachers, parents, testers and authorities) in the field of language testing. She states that "a stakeholder approach to assessment has the effect of democratising assessment processes, of improving relationships between those involved, and promoting greater fairness" (1997: 304). As can be seen, the accountability of the language tester must extend to the public being served. Defining public accountability, however, is a fairly easy task; ensuring accountability to the public less so. Public accountability starts with transparency, of being aware of the kind of information that is made available (Bovens 2005: 10).

In the case of TALPS, the websites and the pamphlets distributed to students will go a long way in ensuring that users are provided with information regarding the

test. Since that information should be available not only to users, but also to the larger public, the challenge is to translate the technical concepts that are embodied in the test and its assessment procedures into more readily accessible, non-specialist language, while at the same time relating their theoretical meaning to real or perceived social and political concerns. Test practices in South Africa are often examined more closely in radio interviews, newspaper reports and interviews, or formal or informal talks. While there is unfortunately no comprehensive record of these, the 'News' tab on the ICELDA website, which dates back to September 2011, nonetheless provides insight into some of the public appearances by ICELDA officials, or of media coverage of test-related issues. In academic and non-academic settings, however, public explanations of how test results can be used must be quite open about both the value and the limitations of language tests, for example that tests are highly useful to identify risk, but still cannot predict everything. Language tests lose their predictive value of future performance, for instance, with every subsequent year of a student's study at university (Visser and Hanslo 2005). Humility about what a test result can tell us, and what it cannot tell us, relates directly to openness about the limited scope of such assessment (Weideman 2014: 8).

If we wish to demonstrate further the required care and concern for others (Weideman 2009: 235) in our test designs, we should also look beyond the assessment, to what happens after the test, and even after the students have completed their studies (Kearns 1998: 140). The reality is that testing the academic literacy of students but doing nothing to help them subsequently may be considered a futile exercise. Issues of accountability dictate that if we test students, we should do something to help them improve the ability that has been measured, if that is indicated by the test results. The responsibility of ethical language testers extends into the teaching that follows, a point that we shall return to below, in the discussion of one institutional arrangement for this. Decisions related to possible interventions are therefore not made in isolation. Rather, these concerns should be uppermost in the minds of test designers, and from an early stage in the design process. The earlier these concerns are acknowledged, even as early on as the initial conceptualisation of the test and its use, the more likely they are to be productively addressed.

4 The Use of the Test Results

One of the key considerations of the test designers when designing a test of this nature is the question of how the results of the test will be used. In a responsible conception of what test design and use involves, there is a shift in responsibility from those who use tests to the test designers themselves. Once the need for TALPS had been established, the next important consideration was, therefore, the interpretation and use of the results of the test.

The use of tests to deny access has been well documented in the literature on language testing (see Fulcher and Davidson 2007; Shohamy 2001). If the focus in education and testing should be on granting rather than denying access (Fulcher and

Davidson 2007: 412), a test like TALPS can be used to do exactly that – facilitate access. As is clear from the introduction, without an intervention nurturing the development of academic literacy that follows the administration of the test, many students may not successfully complete their studies. In discussing the SATAP (The Standardised Assessment Test for Selection and Placement), Scholtz and Allen-lle (2007) observe that an academic literacy test is essential in providing "insight into the intellectual profile and academic readiness of students" and that subsequent

> interventions have positive and financial implications: the individual becomes economically productive, it improves through-put rates and subsidies for institutions, and contributes to economic advancement in South Africa. (Scholtz and Allen-lle 2007: 921)

It is clear that the negative social and other consequences of language assessments that have in the past affected such tests (as discussed by McNamara and Roever 2006: 149 f. under the rubric of "Language tests and social identity", as well as by Du Plessis 2012: 122–124), have been mitigated in tests like TALL and TALPS. These have purposely been designed to assist rather than disadvantage the test taker. The test developers of TALPS insist that should users choose to use the test for access – gaining entry or admission to postgraduate study - rather than placement (post-admission referral or compulsion to enrol for an academic literacy intervention), they should use at least three other criteria or instruments to measure students' abilities rather than rely solely on TALPS. The results of the test should be used with care, since language can never predict all of a candidate's future performance. This is in keeping with AERA's (1999: 146) advice that, in educational settings, a decision or characterisation that will have a major impact on a student should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision (AERA 1999: 146). A mix of information on which to base access decisions has therefore been proposed, with a weighting of 60 % to prior academic performance, and, in line with other findings by South African researchers of fairly low correlations between academic literacy and academic performance (cf. Maher 2011: 33, and the discussion of such investigations), not more than 10–20 % to language ability, with one possible exception:

> This is where the ability is so low (usually in the lowest 7½ % of testees) that it raises ethical questions about allowing those in who so obviously fall short of requirements that they will waste their time and resources on a hopeless venture. (Weideman 2010, Personal communication)

The test developers of TALPS have attempted to consider, from as early as the design stage, to what use the results of the test would be put. This concern with the consequences of the use of the test and its results to make judgements about test takers points directly to issues of responsibility on the part of the test developers. It also links to the issue of the responsible interpretation of test results. We return below to further strategies for minimising the potentially negative impact of test results. In the next section, however, we first consider the impact of the interpretation of test results.

5 Interpreting the Results of the Test

It is also the responsibility of test designers to stipulate how to interpret the results of a test. Because test results almost always have effects (positive and negative) on test takers, it is imperative that a test is administered and scored according to the test developers' instructions (AERA 1999: 61). Thus, an important consideration on the part of the test designers of TALPS was to provide advice to test users and test takers on how to interpret the results of the test. The concern here was that allocating a 'Pass' or 'Fail', or a test score that was difficult to interpret, would stigmatise the students and the test. Instead the test developers of TALL and TALPS use a scale that does not distinguish between a 'Pass' or 'Fail', but rather indicates the level of risk the student has in relation to language, as well as the measures that should be taken to minimise or eliminate such risk. Table 10.3 presents the scoring scale for TALPS, as well as advice to students on how this should be interpreted in the institutional context of the University of Pretoria:

The decision to use this scale, as pointed out by Du Plessis (2012: 108), has been based on years of research undertaken by the test developers of the TALL and TALPS, and the examination of test and average performance scores obtained at different levels of study (Du Plessis 2012). By making the results available in risk bands, they become more interpretable, and, since only the risk bands (in the 'Interpretation' column) are made public, and not the numerical range of the band or the raw mark, the results are less likely to stigmatise. It is also more meaningful, both for test takers and their supervisors, to know the degree of risk (high, clear, less, etc.), rather than to be given a less easily interpretable raw mark.

This interpretation scale opens up another potential refinement to the administration of TALPS, to be discussed in more detail in the next section. This is that, should the test not be fully reliable (which is always the case), those who are possible borderline cases (as in the narrow band of code 3 cases above) could potentially be required to write a second-chance test.

Table 10.3 Guidelines for interpreting the test scores for the TALPS

Code	Interpretation
Code 1 (0–33 %)	High risk: an academic writing intervention (EOT 300) is compulsory
Code 2 (34–55 %)	Clear risk: EOT 300 is compulsory
Code 3 (56–59 %)	Risk: EOT 300 is compulsory
Code 4 (60–74 %)	Less risk: you do not need to enrol for EOT 300
Code 5 (75+)	Little to no risk: you do not need to enrol for EOT 300

6 The Potential for Subsequent Refinement

This section sets out possible refinements that might be made to enhance the effectiveness of TALPS (and by implication other tests of this nature). It does not yet apply to the current version of TALPS, discussed above, but to designs that still have to be thought through further before being realised in practice. The further possibility of refining TALPS derives from a useful and potentially productive proposal of how to deal effectively with, among other elements, the writing component of the current tests. Pot (2013: 53) found that, in the case of TALPS, "the greatest challenge for students is to present a coherent, well-structured academic argument, and to do so by making use of the appropriate communicative functions used in academic discourse. In essence, students fail to grasp the main concept of presenting an academic argument in written form."

It is interesting to note that what she suggests in the case of postgraduate tests of literacy can both equally and profitably be applied to undergraduate tests, which as a rule do not have a separate writing section, as TALPS does. Her proposal may be less resource-intensive than the addition to the undergraduate tests of a full-blown writing task. It is instructive to note, moreover, where this author's engagement with the proposals she makes derives from. As post-entry tests like TALL, TAG, and TALPS began to be widely used, over time the demand from course designers to derive diagnostic information from them has grown. So the initial goal of her research was to untangle, from the mass of information yielded by the test results, that which could assist subsequent course design. The identification of the benefits to be gained from unlocking the diagnostic information of TALPS is the focus of another report (Pot and Weideman 2015), but we wish to focus here only on a number of proposals she makes in the conclusion of her investigation of the diagnostic information to be gleaned from TALPS - proposals that might nonetheless enhance the design of all the post-entry tests discussed in this volume.

Building on design ideas already used, for example, in post-entry tests developed in Australia and New Zealand, Pot's (2013: 54 ff.) proposal is that the designers consider the introduction of a two-tier test. This would mean splitting the test in two, first testing all candidates sitting for the TALPS on the first seven subsections of the test, all of which are in multiple choice format: Scrambled text, Interpreting graphs and visual information, Academic vocabulary, Text types, Understanding texts, Grammar and text relations, and Text editing. Subsequently, should candidates have scored below the cut-off point (currently at 60 %) for this first test, which is an indication of risk, or if they have scored low on the two subtests (Sects. 6 and 7, Grammar and text relations and Text editing) that have in the past shown very high correlations with the writing section, or if they are borderline cases identified through empirical analyses of potential misclassifications related to the reliability of the test (Van der Slik and Weideman 2005: 28), they are given a second opportunity to have their ability to handle academic discourse assessed. In the case of the risk bands outlined in the previous section (Table 10.3), this might be those in the code 3 category.

In a second-chance test, test takers would be given a writing assessment that is at least twice as long as the current 300-word format, i.e. between 500 and 800 words, and perhaps with a further text or texts on the theme of the argument they are expected to write, in addition to the texts that formed part of the first test. The advantages are clear. The essential feature of analytical, academic discourse is distinction-making and the ability to present that coherently in an academic argument. A longer written task would allow them to demonstrate whether they have the ability to structure a comprehensive argument, as well as the ability to acknowledge sources in a conventionally acceptable manner. It would also allow them time to plan properly, a feature that Pot (2013) found may have been missing from the current 90-min test.

This particular refinement to the format of TALPS (and potentially the other tests as well) is mentioned here because the motivation for proposing it derives directly from a consideration of the construct of the test, and what it means subsequently for responsible course design:

> Because distinction-making is at the heart of academic language ability and this study has demonstrated a lack of mainly structural distinction-making in the students' essays, courses can focus on distinction-making as a central theme. (Pot 2013: 58)

7 Providing Writing Support

For those in risk bands that indicate that they should enrol for an academic literacy development course, however, specific provision must be made. At the University of Pretoria, students whose test results show them to be 'at risk' in terms of their academic literacy levels are in the first instance provided with support in the form of a course in academic writing. The intervention that is relevant in this specific instance is the Postgraduate Academic Writing Module, which was developed by the Unit for Academic Literacy (UAL). The test and the course work hand-in-hand. The test is used to determine the academic literacy levels of postgraduate students. Students who are shown to be at risk may be expected by their faculties (larger organisational units, binding together the humanities departments, or engineering, or business sciences, and so forth) at the University of Pretoria to enrol for this module. Having students take the test before the course means that students who are not at risk do not have to sit through a module they may not need. The positive effects are that the test increases awareness among students of their academic literacy levels. In addition, students see the link between the intervention and succeeding in their studies. Table 10.4 highlights the alignment between the sub-tests in TALPS and the tasks students have to complete in the writing course (EOT 300).

There is alignment between assessment and language development: the test and the course are based on the same definition of academic literacy (see Rambiritch 2012).

Table 10.4 Aligning TALPS and EOT 300

Sub-tests in TALPS	What each sub-test tests	Relation to EOT 300
1. Scrambled text	Recognising different parts of a text, forming a cohesive whole	***Task 5, 8***
2. Academic vocabulary	Testing students' knowledge of words used in a specific context	***Theme 1 and 2***
3. Graphic and visual literacy	Interpreting information from a graph, summarising the data, doing numerical computations	***Task 12***
4. Text type	Identifying/classifying different genres/texts types	***Task 4, 6, 7***
5. Comprehension	Reading, classifying and comparing, making inferences, recognising text relations, distinguishing between essential and non-essential information	***Task 1, 2, 9***
6. Grammar and text relations	Sentence construction, word order, vocabulary, punctuation	***Task 8, 10, 13***
7. Editing	Correction of errors in a text	***Task 13***
8. Writing	Argumentative writing, structuring an argument, recognition of sources	***Task 3, 11 and Theme 2***

8 Conclusion

In this narrative of the design and development of TALPS, a key question that we hope to have answered is whether, as test designers, we have succeeded in designing a socially acceptable, fair and responsible test. The need to ask such questions becomes relevant when one works within a framework that incorporates due consideration of the empirical analyses of a test, as well as a concern for the social dimensions of language testing. As fair and responsible test developers it is our objective to ensure that all information about the test, its design, and its use, is freely available to those affected by or interested in its use. It should be the aim of test developers to design tests that are effective, reliable, accessible and transparent, by test developers who are willing to be accountable for their designs. In attempting to satisfy these conditions, the designers of TALPS have intended to ensure that:

- The test is highly reliable, that a systematic validation argument can be proposed for it, and that it is appropriate to be used for the purpose for which it was designed;
- Information about the test is available and accessible to those interested in its design and use;
- The test that can be justified, explained and defended publicly;
- In ensuring transparency they have opened up a dialogue between all those involved in the testing process, as is evidenced in the numerous internet-derived enquiries fielded by ICELDA, the public debate referred to above, as well as in the two perception studies of Butler (2009) and Du Plessis (2012); and
- They have designed a test that is widely perceived, not only by the ever increasing number of users of results, but also by the test takers, to have positive effects;

Being committed to the test takers they serve, and ensuring that their responsibility does not end with a score on a sheet, provide starting points for test developers who wish to design language tests responsibly. It is pleasing to note, in the present case, that the test is followed in almost every case we know of by effective teaching and learning focused on developing those academic literacy abilities that may have put these students at risk of either not completing their studies or not completing their studies in the required time.

Though in having good intentions, designers' rhetoric may well outstrip practice, these intentions have provided a starting point for the designers of TALPS, whose endeavours and goals are perhaps best summarized in the observation that

> our designs are done because we demonstrate through them the love we have for others: it derives from the relation between the technical artefact that is our design and the ethical dimension of our life. In a country such as ours, the desperate language needs of both adults and children to achieve a functional literacy that will enable them to function in the economy and partake more fully of its fruits, stands out as possibly the biggest responsibility of applied linguists. (Weideman 2007: 53)

Our argument has been that in designing TALPS there is a conscious striving to satisfy the requirements for it to be considered a socially acceptable, fair and responsible test. Of course, this narrative does not end here but is intended as the beginning of many more narratives about the test. Ensuring interpretability and the beneficial use of results, the accessibility of information, transparent design and the willingness to defend that design in public – all contribute to responsible test design. It is likely that most of the lessons learned in the development and administration of this postgraduate assessment will spill over into the development or further refinement of other tests of language ability. Tests need to be scrutinised and re-subjected to scrutiny all the time. Each new context in which they are administered calls for further scrutiny and refinement, for determining whether or how the test continues to conform to principles or conditions for responsible test design. As test designers we need to continue to ask questions about our designs, about how trustworthy the measurement is, and how general/specific the trust is that we can place in them.

References

American Educational Research Association (AERA). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Beu, D. S., & Buckley, M. R. (2004). Using accountability to create a more ethical climate. *Human Resource Management Review, 14*, 67–83.

Bovens, M. (2005). Public accountability: A framework for the analysis and assessment of accountability arrangements in the public domain. In E. Ferlie, L. Lynne, & C. Pollitt (Eds.), *The Oxford handbook of public management* (pp. 1–36). Oxford: Oxford University Press.

Boyd, K., & Davies, A. (2002). Doctors' orders for language testers. *Language Testing, 19*(3), 296–322.

Butler, H.G. (2007). A framework for course design in academic writing for tertiary education. PhD thesis, University of Pretoria, Pretoria.
Butler, H. G. (2009). The design of a postgraduate test of academic literacy: Accommodating student and supervisor perceptions. *Southern African Linguistics and Applied Language Studies, 27*(3), 291–300.
Bygate, M. (2004). Some current trends in applied linguistics: Towards a generic view. *AILA Review, 17*, 6–22.
CITO. (2006). *TiaPlus, classical test and item analysis* ©. Arnhem: Cito M. and R. Department.
Davidson, F., & Lynch, B. K. (2002). *Testcraft*. New Haven: Yale University Press.
Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (Eds.). (1999). *Dictionary of language testing*. *Studies in Language Testing, 7*. Cambridge: Cambridge University Press.
Du Plessis, C. (2012). The design, refinement and reception of a test of academic literacy for postgraduate students. MA dissertation, University of the Free State, Bloemfontein.
Frink, D. D., & Klimoski, R. J. (2004). Advancing accountability theory and practice: Introduction to the human resource management review special edition. *Human Resource Management Review, 14*, 1–17.
Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. New York: Routledge.
Geldenhuys, J. (2007). Test efficiency and utility: Longer or shorter tests. *Ensovoort, 11*(2), 71–82.
Hamp-Lyons, L. (2000a). Fairnesses in language testing. In A.J. Kunnan (Ed.), *Fairness and validation in language assessment. Studies in Language Testing, 9*, (pp. 30–34). Cambridge: Cambridge University Press.
Hamp-Lyons, L. (2000b). Social, professional and individual responsibility in language testing. *System, 28*, 579–591.
Hamp-Lyons, L. (2001). Ethics, fairness(es), and developments in language testing. In C. Elder et al. (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies, Studies in Language Testing, 11*, (pp. 222–227). Cambridge: Cambridge University Press.
Inter-Institutional Centre for Language Development and Assessment (ICELDA). (2015). [Online]. Available http://icelda.sun.ac.za. Accessed 7 May 2015.
Kearns, K. P. (1998). Institutional accountability in higher education: A strategic approach. *Public Productivity & Management Review, 22*(2), 140–156.
Kurpius, S. E. R., & Stafford, M. E. (2006). *Testing and measurement: A user-friendly guide*. Thousand Oaks: Sage Publications.
Maher, C. (2011). Academic writing ability and performance of first year university students in South Africa. Research report for the MA dissertation, University of the Witwatersrand, Johannesburg. [Online]. Available http://wiredspace.wits.ac.za/bitstream/. Accessed 20 July 2015.
McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden: Blackwell Publishing.
Mdepa, W., & Tshiwula, L. (2012). Student diversity in South African Higher Education. *Widening Participation and Lifelong Learning, 13*, 19–33.
Naurin, D. (2007). Transparency, publicity, accountability – The missing links. Unpublished paper delivered at the CONNEX-RG 2 workshop on 'Delegation and mechanisms of accountability in the EU', 8–9 March, Uppsala.
Norton, B. (1997). Accountability in language assessment. In C. Clapham & D. Corson (Eds.), *Language testing and assessment: Encyclopaedia of language and education 7* (pp. 323–333). Dordrecht: Kluwer Academic.
Patterson, R., & Weideman, A. (2013). The typicality of academic discourse and its relevance for constructs of academic literacy. *Journal for Language Teaching, 47*(1), 107–123. http://dx.doi.org/10.4314/jlt.v47i1.5.
Pot, A. (2013). Diagnosing academic language ability: An analysis of TALPS. MA dissertation, Rijksuniversiteit Groningen, Groningen.
Pot, A., & Weideman, A. (2015). Diagnosing academic language ability: Insights from an analysis of a postgraduate test of academic literacy. *Language Matters, 46*(1), 22–43. Retrieved from: http://dx.doi.org/10.1080/10228195.2014.986665
Rambiritch, A. (2012). Transparency, accessibility and accountability as regulative conditions for a postgraduate test of academic literacy. PhD thesis, University of the Free State, Bloemfontein.

Rambiritch, A. (2013). Validating the test of academic literacy for postgraduate students (TALPS). *Journal for Language Teaching, 47*(1), 175–193.
Rea-Dickins, P. (1997). So, why do we need relationships with stakeholders in language testing? A view from the U.K. *Language Testing, 14*(3), 304–314.
Scholtz, D., & Allen-lle, C. O. K. (2007). Is the SATAP test an indicator of academic preparedness for first year university students? *South African Journal of Higher Education, 21*(7), 919–939.
Schuurman, E. (2005). *The technological world picture and an ethics of responsibility: Struggles in the ethics of technology*. Sioux Center: Dordt College Press.
Second Language Testing Inc. (2013). *Pilot testing and field testing.* [Online]. Available: http://2lti.com/test-development/pilot-testing-and-field-testing/
Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? Are they fair? *Language Testing, 14*(3), 340–349.
Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. London: Longman.
Shohamy, E. (2008). Language policy and language assessment: The relationship. *Current Issues in Language Planning, 9*(3), 363–373.
Sinclair, A. (1995). The chameleon of accountability: Forms and discourses. *Accounting, Organisations and Society, 20*(2/3), 219–237.
Van der Slik, F., & Weideman, A. (2005). The refinement of a test of academic literacy. *Per Linguam, 21*(1), 23–35.
Van Dyk, T., & Weideman, A. (2004). Finding the right measure: From blueprint to specification to item type. *SAALT Journal for Language Teaching, 38*(1), 15–24.
Visser, A. J., & Hanslo, M. (2005). Approaches to predictive studies: Possibilities and challenges. *South African Journal of Higher Education, 19*(6), 160–1176.
Weideman, A. (2006). Transparency and accountability in applied linguistics. *Southern African Linguistics and Applied Language Studies, 24*(1), 71–86.
Weideman, A. (2007). A responsible agenda for applied linguistics: Confessions of a philosopher. *Per Linguam, 23*(2), 29–53.
Weideman, A. (2009). Constitutive and regulative conditions for the assessment of academic literacy. *South African Linguistics and Applied Language Studies, 27*(3), 235–251.
Weideman, A. (2014). Innovation and reciprocity in applied linguistics. *Literator, 35*(1), 1–10. [Online]. Available doi: http://dx.doi.org/10.4102/lit.v.35i1.1074.
Weideman, A., & Van Dyk, T. (Eds.). (2014). *Academic literacy: Test your competence*. Potchefstroom: Inter-Institutional Centre for Language Development and Assessment (ICELDA).
Weideman, A., Patterson, R., & Pot, A. (2016). Construct refinement in tests of academic literacy. In J. Read (Ed), *Post-admission language assessment of university students*. Dordrecht: Springer.
Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Houndmills: Palgrave Macmillan.

Part V
Conclusion

Chapter 11
Reflecting on the Contribution of Post-Admission Assessments

John Read

Abstract This chapter examines a number of issues arising from the earlier contributions to this volume. It considers the decision by a university about whether to introduce a post-admission language assessment in terms of the positive and negative messages such a decision may convey, as well as the costs versus the benefits. There is some discussion of the need to develop professional communication skills as attributes to enhance the employability of graduates and how such skills can be fostered, along with the development of academic literacy in the disciplines, through various forms of collaboration between English language specialists and academic teaching staff. Finally, it explores ideas related to the concept of English as a lingua franca and what implications they may have for the assessment of university students from different language backgrounds.

Keywords Post-entry language assessment • English language standards in higher education • Professional communication skills • Graduate attributes • Development of academic language skills • Academic literacy instruction • English as a Lingua Franca (ELF)

As specialists in the field, the authors of this volume have naturally focused on the design and delivery of the assessment programme in their respective institutions, with a concern for improving the quality of the measurement of academic language abilities and reporting the results in a meaningful fashion to the various stakeholders. However, this obviously represents a narrow perspective. No matter how good an assessment may be, it will not achieve its desired objectives unless there is strong institutional support at the policy level as well as adequate resourcing – not just for the assessment itself but for effective follow-up action through advising of students and provision of opportunities for academic language development.

J. Read (✉)
School of Cultures, Languages and Linguistics, University of Auckland, Auckland, New Zealand
e-mail: ja.read@auckland.ac.nz

J. Read (ed.), *Post-admission Language Assessment of University Students*, English Language Education 6, DOI 10.1007/978-3-319-39192-2_11

1 Provision for Academic Language Development

In societies like Hong Kong, Oman and South Africa, where a high proportion if not all students entering English-medium universities come from non-English-using backgrounds, the need to further enhance their English language skills is obvious – even if they have had some form of English-medium schooling previously. The language enhancement may be in the form of a foundation programme, compulsory English language courses in the first year of study and beyond, a learning and study skills centre, or (as in the case of Hong Kong) a fourth year added to what has traditionally been a 3-year undergraduate degree.

On the other hand, universities in the major English-speaking countries vary widely in the extent to which they have made provision for the language and learning needs of incoming students, as noted briefly in the Introduction. Universities in the US have a long tradition, going back at least to the 1950s, of freshman composition programmes to develop the academic writing skills of first-year domestic students, and the growth in foreign student numbers from the 1960s led to the parallel development of ESL courses, in the form of both intensive pre-admission programmes and credit courses for degree students. In the UK, the impetus for addressing these issues came initially from the need to ensure that students with English as their second language from Commonwealth countries who were recipients of scholarships and study awards had adequate proficiency in academic English to benefit from their studies in Britain, and summer pre-sessional courses have become an institution in British universities, serving the much broader range of international students who are now admitted. In other English-speaking countries, it has been the liberalising of immigration regulations to allow the recruitment of fee-paying international students which has led to a variety of pre- and post-admission programmes to enhance their academic English skills. The same liberalisation has seen an influx of immigrant families with children who work their way as "English language learners" through the school system to higher education without necessarily acquiring full proficiency in academic English. For such students and for many other domestic students who are challenged by the demands of academic literacy at the tertiary level, there are learning centres offering short courses, workshops, peer tutoring, individual consultations, online resources and so on.

Thus, in a variety of ways universities in the English-speaking countries already offer study support and opportunities for academic language enrichment to their students, at least on a voluntary basis. A proposal to introduce a post-admission language assessment represents a significant further step by seeking to identify students who would benefit from – or perhaps have an obvious need to access – such services in meeting the language demands of their studies. This then leads to the question of whether the assessment and any follow-up action on the student's part should be voluntary or mandatory. It also raises the issue of whether the language and literacy needs revealed by the assessment results may be greater than can be accommodated within existing provisions, meaning that substantial additional funding may be required.

1.1 External and Internal Pressures

In the cases we have seen in this book, some universities are subject to external pressures to address these matters. The controversy over English language standards in Australian universities has already been discussed in the Introduction. In 2012, the Tertiary Education Quality and Standards Agency (TEQSA) announced that its audits of universities in Australia would include comprehensive quality assessments of English language proficiency provisions (Lane 2012). However, a change of government and vigorous lobbying by tertiary institutions asserting that such assessments imposed onerous demands on them led to a ministerial decision that TEQSA would abandon this approach in favour of simply ensuring that minimum standards were being met (Lane 2014a). In the most recent version of the Higher Education Standards Framework, the statutory basis for TEQSA audits, there is just a single explicit reference to English language standards, right at the beginning of the document:

> **1 Student Participation and Attainment**
> **1.1 Admission**
> 1. Admissions policies, requirements and procedures are documented, are applied fairly and consistently, and are designed to ensure that admitted students have the academic preparation and proficiency in English needed to participate in their intended study, and no known limitations that would be expected to impede their progression and completion. (Australian Government 2015)

The change in TEQSA's role was seen as reducing the pressure on tertiary institutions to take specific initiatives such as implementing a post-entry language assessment (PELA), and some such moves at particular universities stalled as a result. Although it is generally recognised that the English language needs of students should be addressed, there is ongoing debate about the most suitable strategy for ensuring that universities take this responsibility seriously (Lane 2014b).

Another kind of external pressure featured in Chap. 6 (this volume). The Oral English Proficiency Test (OEPT) at Purdue University is one example of an assessment mandated by legislation in US states to ensure that prospective International Teaching Assistants (ITAs) have sufficient oral proficiency in English to be able to perform their role as instructors in undergraduate courses. This of course is a somewhat different concern from that of most other post-admission assessments, where the issue is whether the test-takers can cope with the language and literacy demands of their own studies.

In contrast to these cases of external motivation, other post-admission assessments have resulted from internal pressure, in the form of a growing recognition among senior management and academic staff that there were unmet language needs in their linguistically diverse student bodies which could no longer be ignored, particularly in the face of evidence of students dropping out of their first year of study as a result of language-related difficulties. This applies to the original moves towards a PELA at the University of Melbourne (Chap. 2, this volume; see also Elder and Read 2015) in the 1990s, as well as the introduction of the Diagnostic

English Language Needs Assessment (DELNA) at the University of Auckland (Chap. 6, this volume; see also Read 2015b) and what has evolved as the diagnostic assessment procedure for engineering students at Carleton University (Chap. 3, this volume).

2 The Decision to Introduce a Post-Admission Assessment

For universities which are considering the introduction of a post-admission assessment, there are numerous issues to work through. Several useful sources are available to guide institutions in making decisions about whether to introduce a post-admission assessment – preferably in conjunction with a broader strategy to address language and literacy issues among their students – and, if so, how to implement the programme successfully. These sources draw particularly on the experiences of Australian universities with what they call post-entry (or sometimes post-enrolment) language assessments (PELAs), which have grown out of a specific social, educational and political environment over the last 10 years, as explained in the Introduction. However, much of the Australian experience can be applied more widely, in English-speaking countries if not in EMI universities elsewhere.

- The Degrees of Proficiency website (www.degreesofproficiency.aall.org.au) developed from a federally funded project conducted by Katie Dunworth and her colleagues (2013) to survey PELA initiatives in Australian universities and identify the issues faced by the institutions in maintaining English language standards. The website includes a database of existing PELAs and university language policies, links to a range of source materials and other sites, some case studies of programmes at specific universities, and advice on how to implement post-entry assessments as part of a broader strategy for English language development
- In his book on *Standards of English in higher education*, Murray (2016) devotes a chapter to a discussion of the challenges and risks for a university in introducing a PELA. The book builds on Murray's experiences at an Australian university, which provides a case study for a later chapter in the book, but it is also informed by his knowledge of the situation of universities in the UK.
- In a similar vein, Read (2015a) has a chapter outlining "The case for introducing a post-entry assessment", which also considers the pros and cons of such a decision, as well as alternative ways for a university to address students' language and literacy needs.

2.1 Positive and Negative Messages

In his analysis of the advantages and disadvantages of a PELA, Murray (2016, pp. 122–128) gives some emphasis to the kind of messages which are conveyed by using this type of assessment. On the positive side, a PELA can signal to various

stakeholders a commitment on the part of the university to be responsive to the English language needs of incoming students by identifying those at risk of poor academic performance at an early stage. Potentially, it enhances the reputation of the institution if it is seen to be fulfilling its duty of care to the students. Assuming that students being admitted to the university through various pathways all take the same assessment, the PELA also provides an equitable basis for allocating English language tutoring and other specialist resources to the students who are most at risk. Thus, if the commitment is genuinely made, it reflects well on the institution in meeting its ethical responsibilities to a linguistically diverse student body.

On the other hand, Murray points out that the messages may be negative. He reports from his observations that university senior management are very cautious about any form of PELA because, first, it may indicate that the university has lowered its standards by accepting students who are linguistically weak, and, secondly, it may put off potential students when they learn that they face an additional hurdle after meeting the normal admission requirements, and in particular after "passing" IELTS or TOEFL. Murray suggests how a university can be proactive in countering such concerns through the way that it presents the rationale for the PELA to external stakeholders. In addition, he recommends that the assessment should be conducted in a low-key fashion through faculties and departments, rather than as a high-profile, mandatory and centrally administered programme which is more likely to attract criticism and complaint from students.

This last point is taken up by Read (2008), in his discussion of how the Diagnostic English Language Needs Assessment (DELNA) has been promoted internally at the University of Auckland. Read draws on Read and Chapelle's (2001) concept of test presentation, defined as "a series of steps, taken as part of the process of developing and implementing the test, to influence its impact in a positive direction" (p. 185). In the early years of administering DELNA, before it became mandatory, mature students and others with no recent history of study in New Zealand would receive a letter from the Admissions office inviting them to take the assessment and emphasising its potential value as a diagnosis of their academic language ability. To reach a broader range of students DELNA staff speak to students at Orientation and other events about the benefits of the assessment; there are posters, bookmarks and webpages which offer a "free health check of your academic English language skills" and feature slogans such as "Increase your chance of success" and "Students say DELNA is time well spent". Every effort has been made to embed the assessment as just one more task that first-year students need to complete in order to enter the university.

Similarly, there are ongoing efforts to inform academic and professional staff at Auckland about the programme. The main vehicle is the DELNA Reference Group, composed of representatives from all the faculties and relevant service units around the university, which meets twice a year to discuss policy issues, monitor student compliance with the DELNA requirements, and provide a channel of communication to staff. In addition, the DELNA Manager is active in briefing and liaising with key staff members on an individual basis, and there is an FAQ document which

addresses common questions and concerns. Through all these means, the university seeks to ensure that the purpose of the assessment is understood, and that students take advantage of the opportunities it offers.

2.2 Costs and Benefits

The costs of introducing a post-admission assessment often weigh heavily on those charged with making the decision. The direct expenses of developing the instruments and administering them are the most obvious ones, but then there are also the associated costs of enhanced provision for English language development to cater for the needs of the students who perform poorly on the assessment. As Murray puts it, "to deprive these students of such opportunities [for development] would undermine the credibility of the institution and its English language initiative, and call into question its clarity of thinking and the commitment it has to those students and to the English agenda more generally" (2016, p. 127).

Based on her survey of Australian universities, Dunworth (2009) found a number of concerns about the resources associated with a PELA. Many of her respondents were worried that there would not be adequate funding to meet the needs revealed by the assessment, especially if the PELA itself consumed a disproportionate amount of the budget for student services. This was more of an issue when the assessment was designed for a particular School or Faculty, which would obviously have a more limited funding base than the central university budget. There was a tendency for university managers to underestimate the resources required to implement a good-quality assessment programme as well as the need to plan ahead for appropriate follow-up strategies.

An interesting perspective on the relative costs and benefits of a post-admission assessment is found in Chap. 3, where the Dean of Engineering and Design at Carleton University in Canada is quoted as saying, with reference to three students who remained in the undergraduate programme rather than dropping out, "Retaining even two students pays for the expense of the entire academic assessment procedure" (this volume, p. xx).

Along the same lines the Deputy Vice-Chancellor (Academic) at the University of Auckland, who has management responsibility for the University's Diagnostic English Language Needs Assessment (DELNA), reasons this way:

> If one looks at the high level figures, it is easy to see the picture. You can divide the DELNA budget by the funding which the university receives for each fulltime student to get an idea of how many students we need to retain as a result of DELNA impact to protect our revenue. This deals with future revenue lost by the university, and it amounts to around 20 students. There is also the matter of the past wasted financial costs to students and the government (50/50 to each party) when students withdraw or are excluded for reasons that can be traced to their inadequate academic English; to these costs can be added the income foregone by

> students when they have been attending university to little purpose rather than working—probably $20,000 per student. Then there are all the non-financial costs—angst, frustrated expectations and so on. (John Morrow, personal communication, 8 March 2016)

This quote refers specifically to the costs of the assessment, but the same line of argument can be extended to the funding needed for a programme of academic language development, much of which was already in place at the time that DELNA was introduced.

3 Extending the Scope of Academic Language Development

One criticism of post-admission assessments is that by definition they are administered when students first arrive on campus and, as we have seen in the chapters of this volume, follow-up language development programmes are concentrated in the first year of study. The implicit assumption is that early intervention is the best strategy (and perhaps all that is needed) for addressing the students' needs. However, it is worth recalling from the Introduction that Birrell's (2006) paper which prompted public debate in Australia about the English proficiency of international students was concerned with the evidence that they were graduating with inadequate command of the language to be employable in that country, rather than whether they could cope with the language demands of their academic studies.

3.1 Professional Communication Skills

A logical response to Birrell's work, then, would be to determine whether students have the language skills they need for future employment at the time they complete their undergraduate degree. This is consistent with the current practice in Australian, New Zealand and British universities of specifying generic graduate attributes, which are defined in this widely quoted statement as:

> the qualities, skills and understandings a university community agrees its students should develop during their time with the institution. These attributes include but go beyond the disciplinary expertise or technical knowledge that has traditionally formed the core of most university courses (Bowden et al. 2000, cited in University of Edinburgh 2011).

In the policy documents of particular universities in English-speaking countries, language tends to figure under the guise of "effective communication", as in these examples:

University of Melbourne:

> Melbourne graduates … can apply knowledge, information and research skills to complex problems in a range of contexts and are effective oral and written communicators. (http://msl.unimelb.edu.au/teaching-learning)

University of Sydney:

5. Communication

   Graduates of the University will recognise and value communication as a tool for negotiating and creating new understanding, interacting with others, and furthering their own learning.

- use oral, written, and visual communication to further their own learning
- make effective use of oral, written and visual means to critique, negotiate, create and communicate understanding
- use communication as a tool for interacting and relating to others (http://www.itl.usyd.edu.au/graduateAttributes/policy_framework.pdf)

However, as with other graduate attributes, there is a lack of university-wide strategies to determine whether graduating students have acquired such communication skills, except through the assessment of the courses they have taken for their degree. As Arkoudis & Kelly put it,

> institutional graduate attribute statements that refer to the communication skills of graduates are merely claims until evidenced. Institutional leaders need to be able to point to evidence demonstrating that the oral and written communication skills of their students are developed, assessed, monitored and measured through the duration of a qualification. (2016, p. 6)

They go on to note the need for research to articulate exit standards and to produce an explicit framework which could guide academic staff to develop the relevant skills through the teaching of their courses.

As a step in this direction, Murray (2010, 2016) proposes that the construct of English language proficiency for university study should be expanded to include professional communication skills, of the kind that students will require both for work placements and practicums during their studies and in order to satisfy the expectations of future employers and professional registration bodies once they graduate. Murray identifies these skills as follows:

- Intercultural competence
- A cultural relativistic orientation
- Interpersonal skills
- Conversancy in the discourses and behaviours associated with particular domains
- Non-verbal communication skills
- Group and leadership skills

The one language testing project which has sought to produce a measure of at least some of these skills is the Graduating Students' Language Proficiency Assessment (GSLPA), developed in the 1990s at Hong Kong Polytechnic University (PolyU), with funding from the University Grants Committee (UGC) in Hong Kong (Qian 2007). It is a task-based test of professional writing and speaking skills designed in consultation with business leaders in Hong Kong. Although the test has been administered to PolyU students since 1999 (see http://gslpa.polyu.edu.hk/eng/web/), it was not accepted by the other Hong Kong universities and, as an alternative,

the UGC ran a scheme from 2002 to 2013 to pay the fee for students to take the Academic Module of IELTS on a voluntary basis when they were completing their degree. Two Australian universities (the University of Queensland and Griffith University) have adopted a similar policy of subsidising the IELTS test fee as a service to their graduating international students (Humphreys and Mousavi 2010). While this strategy provides the students with a broad, internationally recognised assessment of their academic language proficiency at the time of graduation, it can scarcely be regarded as a valid measure of their professional communication skills. Indeed, O'Loughlin (2008) has questioned the ethics of using IELTS for such a purpose without proper validation.

3.2 Embedded Language Development

A quite different approach involves embedding these skills, along with other aspects of English language development, into the students' degree programmes. This already happens to varying degrees in professional faculties, like Engineering, Business, Medical Sciences and Education, where students need to demonstrate the application of relevant communication skills in order to be registered to practise their chosen profession. The same strategy can in principle be applied to degree programmes across the university. Numerous English language specialists in higher education – notably Arkoudis et al. (2012) in Australia and Wingate (2015) in the United Kingdom – strongly advocate the embedded delivery of academic language development to all students as current best practice. In support of this position, Arkoudis and Kelly cite studies which document "the limitations of communication skills programs which sit outside the disciplinary curricula and are supported by staff who are not recognised by students as disciplinary academics" (2016, p. 4).

This quote highlights the point that academic English programmes are typically delivered as adjuncts to degree courses by tutors with low (and maybe insecure) status within the institution who may not have the relevant knowledge of discourse norms to address issues of academic literacy or professional communication skills within the disciplines. On the other hand, subject lecturers and tutors tend to shy away from dealing with problems with language and genre in their students' writing, claiming a lack of expertise. In their influential study of academic literacies in undergraduate courses in the UK, Lea and Street (1998) reported that tutors could not adequately articulate their understanding of concepts like "critical analysis", "argument" or "clarity". As Murray (2016) puts it, although academic teaching staff have procedural knowledge of academic discourse norms in their discipline, they lack the declarative (or metalinguistic) knowledge needed to give the kind of feedback on student writing that would allow the students to understand how they can better meet the appropriate disciplinary norms.

This suggests that the way forward is to foster more collaboration between learning advisors and English language tutors on the one hand and academic teaching staff on the other. Murray (2016) proposes as a starting point that the practice in

some universities of locating language tutors within particular faculties should be more widely adopted, to give more opportunities for interaction between the two sides. Drawing on their extensive experience as learning advisors at the University of Sydney, Jones et al. (2001) outline four models of collaboration in the development of academic writing skills. At the most basic level, there is a "weak adjunct" model which provides generic tutorials on academic writing outside of class hours. A "strong adjunct" model is delivered in a similar fashion but with a focus on writing genres that are relevant to the students' discipline, such as lab reports or research proposals. Then comes the "integrated model" in which learning advisors give presentations or workshops on discipline-specific aspects of academic literacy during class hours. At the top level, a fully "embedded" model involves a course curriculum with a primary focus on literacy in the discipline, designed collaboratively by learning advisors and the subject lecturers who will actually teach the course.

The integrated and embedded models clearly require a significant ongoing commitment of time and resources by both parties, which is difficult to initiate and even more challenging to sustain. Arkoudis et al. (2012) describe a version of the integrated model which was conducted for one semester in an Architecture course at the University of Melbourne, with promising results, but they acknowledge that the model could not be widely implemented on a regular basis. As alternatives, they discuss ways in which course coordinators can incorporate academic literacy goals into the grading of course assignments and can foster productive interactions among their students through the careful design of group discussions and projects, with the active involvement of English language specialists.

Wingate (2015) makes a strong case for what she calls "inclusive practice" to overcome the limitations of current approaches to academic literacy development. This means applying four principles, which can be summarised as follows:

1. Academic literacy instruction should focus on an understanding of the genres associated with the students' academic subjects, rather than taking the generic approach found in the typical EAP programme.
2. *All* students should have access to this instruction, regardless of their language background. Any language support for non-native speakers should be provided in addition to the academic literacy instruction.
3. The instruction needs to be integrated with the teaching of content subjects so that ideally academic literacy is assessed as part of the subject curriculum.
4. Academic literacy instruction requires collaboration between writing experts and subject experts to develop the curriculum jointly (2015, pp. 128–130).

As a first step, Wingate describes how she and her colleagues at Kings College London have designed and delivered academic literacy workshops for students in four disciplines, but she recognises that substantial cultural and structural changes would be necessary to implement the four principles throughout a whole university. Nevertheless, she argues that longer term trends will force institutions to move in this direction: "market forces such as growing competition for students and expectations by high-fee paying students will increase the need for universities to provide effective support for students … from diverse backgrounds" (2015, p. 162).

Full implementation of Wingate's principles would reduce, if not eliminate, the need for post-admission language assessment – but that prospect seems rather distant at this point.

4 The ELF Perspective

One further perspective to be considered is represented by the term English as a Lingua Franca (ELF). In Chap. 8, Roche et al. have adopted the term to refer to the status of English in the Omani universities in which they conducted their research. At one level, it can be seen as a synonym for English as an International Language (EIL), a relatively neutral description of the current dominance of the language as a means of communication across national and linguistic boundaries, as well as the prime vehicle for globalisation in social, economic, scientific, educational and cultural terms. However, during the last 15 years ELF has come to represent in applied linguistics a more critical perspective on the role of English internationally and, more particularly, the status of native speakers and their brand of English. Non-native users of the language greatly outnumber native speakers on a worldwide basis and a large proportion of daily interactions in the language do not involve native speakers at all. This calls into question the "ownership" of English (Widdowson 1994) and the assumed authority of native speakers as models or arbiters of accuracy and appropriateness in the use of the language.

To substantiate this argument, a large proportion of the ELF research has drawn on spoken language corpora – the Vienna-Oxford International Corpus of English (VOICE) (Seidlhofer 2011), English as a Lingua Franca in Academic Settings (ELFA) (Mauranen 2012) and the Asian Corpus of English (ACE) (Kirkpatrick 2010) – featuring mostly well-educated non-native speakers of English from different countries communicating with each other. Apart from providing descriptions of recurring grammatical and lexical features in these oral interactions, researchers have highlighted communicative strategies that anticipate or repair potential breakdowns in mutual comprehension, putting forth the argument that non-native users of English are more adept at dealing with such situations than native speakers are.

One of the most prominent ELF advocates, Jennifer Jenkins (2013), has turned her attention in a recent book to English-medium instruction (EMI) in universities, both those in the traditionally English-speaking countries and the increasing number of institutions, particularly in Europe, the Middle East, and East and Southeast Asia, which offer degree programmes in English as well as their national language. From an analysis of university websites and a questionnaire survey of 166 academics, Jenkins concluded that institutional claims to the status of an "international university" for the most part did not extend to any recognition of the role of English as a lingua franca, or any corresponding challenge to the dominance of native speaker norms. Most of the questionnaire respondents apparently took it for granted that the best guarantee of maintaining high academic standards was to expect second language users to adhere (or at least aspire) to native speaker English. However,

they also acknowledged that the level of support offered by their university to non-native English speakers was inadequate, with consequent negative effects on students' confidence in their ability to meet the standards.

The latter view received support in a series of "conversations" Jenkins (2013) conducted at a UK university with international postgraduate students, who expressed frustration at the lack of understanding among their supervisors, lecturers and native-speaking peers concerning the linguistic challenges they faced in undertaking their studies. This included an excessive concern among supervisors with spelling, grammar and other surface features as the basis for judging the quality of the students' work – often with the rationale that a high level of linguistic accuracy was required for publication in an academic journal.

4.1 ELF and International Proficiency Tests

Jenkins (2013; see also Jenkins 2006a; Jenkins and Leung 2014) is particularly critical of the role of the international English proficiency tests (IELTS, TOEFL, Pearson Test of English (PTE)) in their gatekeeping role for entry to EMI degree programmes. She and others (e.g., Canagarajah 2006; Clyne and Sharifian 2008; Lowenberg 2002) argue that these and other tests of English for academic purposes serve to perpetuate the dominance of standard native-speaker English, to the detriment of ELF users, by requiring a high degree of linguistic accuracy, by associating an advanced level of proficiency with facility in idiomatic expression, and by not assessing the intercultural negotiating skills which are a key component of communication in English across linguistic boundaries, according to the ELF research. These criticisms have been largely articulated by scholars with no background in language assessment, although Shohamy (2006) and McNamara (2011) have also lent some support to the cause.

Several language testers (Elder and Davies 2006; Elder and Harding 2008; Taylor 2006) have sought to respond to the criticisms from a position of openness to the ideas behind ELF. Their responses have been along two lines. On the one hand, they have discussed the constraints on the design and development of innovative tests which might more adequately represent the use of English as a lingua franca, if the tests were to be used to make high-stakes decisions about students. On the other hand, these authors have argued that the critics have not recognised ways in which, under the influence of the communicative approach to language assessment, contemporary English proficiency tests have moved away from a focus on native-speaker grammatical and lexical norms towards assessing a broader range of communicative abilities, including those documented in ELF research. The replies from the ELF critics to these statements (Jenkins 2006b; Jenkins and Leung 2014) have been disappointingly dismissive, reflecting an apparent disinclination to engage in constructive debate about the issues.

This is not to say that the international proficiency tests are above criticism. Language testers can certainly point to ways in which these testing programmes

under-represent the construct of academic language proficiency and narrow the horizons of students who undertake intensive test preparation at the expense of a broader development of their academic language and literacy skills. IELTS and TOEFL are prime exemplars of what Spolsky (1995, 2008) has labelled "industrial language testing", being administered to around two million candidates each at thousands of test centres around the world. This means that there are huge resources invested, not just in the tests themselves but in the associated test preparation industry, and as a consequence it is a major undertaking to make any substantive changes to the tests of the kind that ELF advocates would like to see.

4.2 ELF and Post-Admission Assessments

This brings us back to the role of post-admission assessments. As things stand at present, and for the foreseeable future, such assessments cannot realistically replace tests like IELTS, TOEFL or PTE for pre-admission screening of international students because most universities take it for granted that a secure, reliable test of this kind is an essential tool in the admissions process and, in the cases of Australia and the United Kingdom, the immigration authorities specify a minimum score on a recognised English test as a prerequisite for the issuing of a student visa. However, post-admission assessments developed for particular universities can complement the major tests by representing flexible responses to local circumstances and to changing ideas about appropriate forms of assessment, such as those associated with ELF.

Perhaps the most revealing finding from Jenkins' (2013) surveys was the extent to which academics in the UK and in EMI institutions elsewhere defined academic standards in traditional terms which favoured native-speaking students, and many appeared insensitive to ways in which they could modify their teaching and supervisory practices to accommodate international students, without "dumbing down" the curriculum. The introduction of a post-admission assessment will do nothing in itself to shift such attitudes. If an assessment is implemented in such an environment, it may basically perpetuate a deficit model of students' language needs, which places the onus squarely on them (with whatever language support is available to them) to "improve their English", rather than being part of a broader commitment to the promotion of high standards of academic literacy for all students, regardless of their language background.

One issue here is whether incoming students for whom English is an additional language should be considered to have the status of "learners" of English, rather than non-native "users" of the language who need to enhance their academic literacy skills in the same way that native-speaking students do. Most of the ELF literature focuses on non-native users who are already highly proficient in the language, so that the distinctive linguistic features in their speech represent relatively superficial aspects of what is actually a high level of competence in a standard variety of English. A good proportion of international doctoral students potentially

fall into this category, particularly if they have already had the experience of using English for purposes like presenting their work at conferences or writing for publication in English. On the other hand, a diagnostic assessment may reveal that such students read very slowly, lack non-technical vocabulary knowledge, have difficulty in composing cohesive and intelligible paragraphs, and are hampered in other ways by limited linguistic competence. This makes it more arguable whether such students should be considered proficient users of the language.

A similar kind of issue arises with first-year undergraduates in English-speaking countries matriculating from the secondary school system there. Apart from international students who complete 2 or 3 years of secondary education to prepare for university admission, domestic students cover a wide spectrum of language backgrounds which make it increasingly problematic to distinguish non-native users from native speakers in terms of the language and literacy skills required for academic study. In the United States English language learners from migrant families have been labelled Generation 1.5 (Harklau et al. 1999; Roberge et al. 2009) and are recognised as often being in an uncomfortable in-between space where they have not integrated adequately into the host society, culture and education system. Linguistically, they may have acquired native-like oral communication skills, but they lack the prerequisite knowledge of the language system on which to develop good academic reading and writing skills. Such considerations strengthen the case for administering a post-admission assessment to all incoming students, whatever their language background; this is the position of the University of Auckland with DELNA, but not many universities have been able to adopt a comprehensive policy of this kind.

At the same time, there are challenging questions about how to design a post-admission assessment to cater for the diverse backgrounds of students across the native – non-native spectrum. It seems that the ELF literature has little to offer at this point towards the definition of an alternative construct of academic language ability which avoids reference to standard native-speaker norms and provides the basis for a practicable assessment design. The work of Weideman and his colleagues in South Africa, on defining and assessing the construct of academic literacy, as reported in Chaps. 9 and 10, represents one stimulating model of test design, but others are needed, especially if post-admission assessments are to operationalise an academic literacies construct which takes account of the discourse norms in particular academic disciplines, as analysed by scholars such as Swales (1990), Hyland (2000, 2008), and Nesi and Gardner (2012). At the moment the closest we have to a well-documented assessment procedure of this type is the University of Sydney's Measuring the Academic Skills of University Students (MASUS) (Bonanno and Jones 2007), as noted in the Introduction.

Nevertheless, the chapters of this volume show what can be achieved in a variety of English-medium universities to assess the academic language ability of incoming students at the time of admission, as a prelude to the delivery of effective programmes for language and literacy development. It is important to acknowledge that all of the institutions represented here have been able to draw on their own applied linguists and language testers in designing their assessments. As Murray noted in

identifying universities "at the vanguard" of PELA provision in Australia and New Zealand, "It is certainly not coincidental that a number of these boast resident expertise in testing" (2016, p. 121). The converse is that institutions lacking such capability may implement assessments which do not meet professional standards. However, by means of publications and conference presentations, as well as consultancies and licensing arrangements, the expertise is being more widely shared, and we hope that this book will contribute significantly to that process of dissemination.

References

Arkoudis, S., & Kelly, P. (2016). *Shifting the narrative: International students and communication skills in higher education* (IERN Research Digest, 8). International Education Association of Australia. Retrieved March 1, 2016, from: www.ieaa.org.au/documents/item/664

Arkoudis, S., Baik, C., & Richardson, S. (2012). *English language standards in higher education*. Camberwell: ACER Press.

Australian Government. (2015). *Higher education standards framework (threshold standards) 2015*. Retrieved March 7, 2016, from: https://www.legislation.gov.au/Details/F2015L01639

Birrell, B. (2006). Implications of low English standards among overseas students at Australian universities. *People and Place, 14*(4), 53–64. Melbourne: Centre for Population and Urban Research, Monash University.

Bonanno, H., & Jones, J. (2007). *The MASUS procedure: Measuring the academic skills of university students. A resource document*. Sydney: Learning Centre, University of Sydney. http://sydney.edu.au/stuserv/documents/learning_centre/MASUS.pdf

Canagarajah, A. S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an international language. *Language Assessment Quarterly, 3*(3), 229–242.

Clyne, M., & Sharifian, F. (2008). English as an international language: Challenges and possibilities. *Australian Review of Applied Linguistics, 31*(3), 28.1–28.16.

Dunworth, K. (2009). An investigation into post-entry English language assessment in Australian universities. *Journal of Academic Language and Learning, 3*(1), 1–13.

Dunworth, K., Drury, H., Kralik, C., Moore, T., & Mulligan, D. (2013). *Degrees of proficiency: Building a strategic approach to university students' English language assessment and development*. Sydney: Australian Government Office for Learning and Teaching. Retrieved February 24, 2016, from: www.olt.gov.au/project-degrees-proficiency-building-strategic-approach-university-studentsapos-english-language-ass

Elder, C., & Davies, A. (2006). Assessing English as a lingua franca. *Annual Review of Applied Linguistics, 26*, 282–304.

Elder, C., & Harding, L. (2008). Language testing and English as an international language: Constraints and contributions. *Australian Review of Applied Linguistics, 31*(3), 34.1–34.11.

Elder, C., & Read, J. (2015). Post-entry language assessments in Australia. In J. Read (Ed.), *Assessing English proficiency for university study* (pp. 25–39). Basingstoke: Palgrave Macmillan.

Harklau, L., Losey, K. M., & Siegal, M. (Eds.). (1999). *Generation 1.5 meets college composition: Issues in the teaching of writing to U.S.-educated learners of ESL*. Mahwah: Lawrence Erlbaum.

Humphreys, P., & Mousavi, A. (2010). Exit testing: A whole-of-university approach. *Language Education in Asia, 1*, 8–22.

Hyland, K. (2000). *Disciplinary discourses: Social interactions in academic writing*. Harlow: Longman.

Hyland, K. (2008). Genre and academic writing in the disciplines. *Language Teaching, 41*(4), 543–562.
Jenkins, J. (2006a). The spread of EIL: A testing time for testers. *ELT Journal, 60*(1), 42–50.
Jenkins, J. (2006b). The times they are (very slowly) a-changin'. *ELT Journal, 60*(1), 61–62.
Jenkins, J. (2013). *English as a lingua franca in the international university: The politics of academic English language policy*. London: Routledge.
Jenkins, J., & Leung, C. (2014). English as a lingua franca. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1–10). Chichester: Wiley. Chap. 95.
Jones, J., Bonanno, H., & Scouller, K. (2001). *Staff and student roles in central and faculty-based learning support: Changing partnerships*. Paper presented at Changing Identities, 2001 National Language and Academic Skills Conference. Retrieved April 17, 2014, from: http://learning.uow.edu.au/LAS2001/selected/jones_1.pdf
Kirkpatrick, A. (2010). *English as a Lingua Franca in ASEAN: A multilingual model*. Hong Kong: Hong Kong University Press.
Lane, B. (2012, August 22). National regulator sharpens focus on English language standards. *The Australian*. Retrieved March 7, 2016, from: www.theaustralian.com.au/higher-education/national-regulator-sharpens-focus-onenglish-language-standards/story-e6frgcjx-1226455260799
Lane, B. (2014a, March 12). English proficiency at risk as TEQSA bows out. *The Australian*, March 12. Retrieved March 7, 2016, from: http://www.theaustralian.com.au/higher-education/english-proficiency-at-risk-as-teqsa-bows-out/story-e6frgcjx-1226851723984
Lane, B. (2014b, August 22). Unis and language experts at odds over English proficiency. *The Australian.* Retrieved March 7, 2016, from: http://www.theaustralian.com.au/higher-education/unis-and-language-experts-at-odds-over-english-proficiency/news-story/d3bc1083caa28eb8924e94b0d40b0928
Lea, M. R., & Street, B. V. (1998). Student writing in higher education: An academic literacies approach. *Studies in Higher Education, 29*, 157–172.
Lowenberg, P. H. (2002). Assessing English proficiency in the expanding circle. *World Englishes, 21*(3), 431–435.
Mauranen, A. (2012). *Exploring ELF: Academic English shaped by non-native speakers*. Cambridge: Cambridge University Press.
McNamara, T. (2011). Managing learning: Authority and language assessment. *Language Teaching, 44*(4), 500–515.
Murray, N. (2010). Considerations in the post-enrolment assessment of English language proficiency: From the Australian context. *Language Assessment Quarterly, 7*(4), 343–358.
Murray, N. (2016). *Standards of English in higher education: Issues, challenges and strategies*. Cambridge: Cambridge University Press.
Nesi, H., & Gardner, S. (2012). *Genres across the disciplines: Student writing in higher education*. Cambridge: Cambridge University Press.
O'Loughlin, K. (2008). The use of IELTS for university selection in Australia: A case study. *IELTS Research Reports, Volume 8* (Report 3). Retrieved March 11, 2016, from: https://www.ielts.org/~/media/research-reports/ielts_rr_volume08_report3.ashx
Qian, D. (2007). Assessing university students: Searching for an English language exit test. *RELC Journal, 38*(1), 18–37.
Read, J. (2008). Identifying academic language needs though diagnostic assessment. *Journal of English for Academic Purposes, 7*(2), 180–190.
Read, J. (2015a). *Assessing English proficiency for university study*. Basingstoke: Palgrave Macmillan.
Read, J. (2015b). The DELNA programme at the University of Auckland. In J. Read (Ed.), *Assessing English proficiency for university study* (pp. 47–69). Basingstoke: Palgrave Macmillan.
Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing, 18*(1), 1–32.

Roberge, M., Siegal, M., & Harklau, L. (Eds.). (2009). *Generation 1.5 in college composition: Teaching academic writing to U.S.-educated learners of ESL*. New York: Routledge.
Seidlhofer, B. (2011). *Understanding English as a Lingua Franca*. Oxford: Oxford University Press.
Shohamy, E. (2006). *Language policy: Hidden agendas and new approaches*. London: Routledge.
Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford: Oxford University Press.
Spolsky, B. (2008). Language assessment in historical and future perspective. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (Language testing and assessment 2nd ed., Vol. 7, pp. 445–454). New York: Springer.
Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.
Taylor, L. (2006). The changing landscape of English: Implications for language assessment. *ELT Journal, 60*(1), 51–60.
University of Edinburgh. (2011). Employability initiative at Edinburgh. Retrieved March 9, 2016, from: http://www.employability.ed.ac.uk/GraduateAttributes.htm
Widdowson, H. G. (1994). The ownership of English. *TESOL Quarterly, 28*(2), 377–389.
Wingate, U. (2015). *Academic literacy and student diversity: The case for inclusive practice*. Bristol: Multilingual Matters.

Index

J. Read (ed.), *Post-admission Language Assessment of University Students*,
English Language Education 6, DOI 10.1007/978-3-319-39192-2