John Wrigglesworth

Pre-sessional listening assessment: Construct, content and graduateness

Introduction

Many pre-sessional course leaders are implementing an English for Specific Academic Purposes (ESAP) focus to their curriculum to increase the value of the provision to disciplinary sub-groups within their cohort of students. However, writing reliable and valid assessments for these sub-groups is a challenge as pre-sessional students have a range of target courses across many disciplines. Perhaps the most difficult language skill to write assessments for is listening (Field, 2011). Developing a range of discipline-specific listenings is beyond the resources of most language centres; it is more efficient to assess a whole cohort with one version of a listening test. This problem is offered as a reason to remain with English for General Academic Purposes (EGAP).

Rethinking the scope of an ESAP pre-sessional's learning objectives may help. A catalyst to this rethinking comes from graduateness research. Research into graduateness has articulated the skills and attributes that all students gain in higher education (Coetzee, 2014). Graduateness frameworks offer a principled way of identifying pre-sessional assessment constructs and specifications that are broader than a specific discipline, but still include them. Doing so allows a pre-sessional course to include 'general' aspects of graduateness in ESAP learning objectives and to assess ESAP differently across language skills.

In this paper, the development of a pre-sessional listening test at a post-92 university in the UK is used to exemplify some of the challenges and compromises made during test development which followed such a rethinking. It suggests how notions of 'graduateness' can widen the scope of the needs analysis, refine ESAP learning objectives and inform manageable, but robust, listening assessments.

A tight spot for pre-sessional assessment

Initial definitions of EAP start with its purpose: 'to help English learners develop the skills they need to study through the medium of English' (Charles & Pecorari, 2016, p. 1). A familiar context for learners to develop these skills is taking a pre-sessional course; these courses exist to prepare students for future academic study, particularly those with an offer that is conditional on demonstrating an improved level of English (de Chazal, 2014, p. 33). Studying in a foreign language is difficult and so 'in order to deliver maximum value to their students', pre-sessional courses need to focus on specialized language (Charles & Pecorari, 2016, p. 8).

For course leaders, developing and assessing bespoke pre-sessional courses is challenging. Maximizing value through the principled differentiation of student experience is resource intensive; the EGAP/ESAP debate shows EAP practitioners responding to the conflicting forces of the practical and the ideal (see de Chazal, 2014, p. 38 and Hyland, 2016, for contrasting perspectives). For some, pre-sessional courses cannot usually meet the discipline-specific needs of all their students (e.g., Bruce, 2011). Those that try require assessments that align with their ESAP learning objectives. However, as Manning shows (2018), the process of developing valid and reliable EGAP assessments involves a considerable investment of time and resources. This investment is multiplied for a diverse cohort of students with every ESAP iteration of a test.

For students, educational assessment is clearly high-stakes. Failure for any student can result in suspending studies or not progressing to a chosen career path. An international student on a UK pre-sessional will already have relocated to a new country, paid tuition and accommodation fees, and committed to considerable additional costs (e.g., visa, NHS health surcharge). They are unlikely to have offers for courses elsewhere and opportunities to take other tests exigently (such as IELTS) are limited.

Higher Educational Institutions (HEIs) who want to recruit and sponsor international students must have Highly Trusted Status (HTS) with UKVI and hold a Tier 4 sponsor license. UKVI carries out audits of Tier 4 sponsor institutions. Their guidance for English language assessment states that, 'As a HEI, you are free to assess English language by any means you see fit, but you must ensure they [the students] are proficient at B2 in all 4 components prior to issuing a (CAS) [Confirmation of Acceptance for Studies] at level 6 or above.' (Home Office, 2019). They go on to caution that, 'We will FULLY scrutinise any in-house English test and document how the test operates. We will examine test papers for evidence of all 4 areas. We will NOT, however, comment on the structure or contents of the test'.

The most straightforward way for a course leader and HEI to ensure continued HTS is to follow best practice in language assessment. In the case of academic listening, the course leader needs to be able to define what academic listening is and defend how they assess it.

Challenges for listening constructs in university contexts

Listening is a complex, and arguably the least researched, skill in EAP (Lynch, 2011).

Rost (2002) surveys the wide-ranging second language acquisition literature on listening and the ways that listening is measured (i.e., operationalizing the construct). Vandergrift (2004) reviews a range of listening pedagogy research and offers a model which integrates strategic metacognitive awareness, bottom-up decoding and strategies for teaching the process of how best to listen. Field (2019) builds on current understanding of what listening is and applies it to language testing theory and the need to test at different levels of proficiency. Within EAP, Bruce (2011, pp. 154–176) focuses on the processes for developing listening skills (particularly top-down and bottom-up processing) and the knowledge that learners draw on when listening (contextual, semantic, syntactic, lexical and phonological); he applies these to the 'key extended monologic listening event' of the university lecture. Alexander, Argent and Spencer (2008, pp. 217–226) discuss issues around the purposes and authenticity of lecture materials, pointing out that identifiable features of authenticity can be controlled to meet the learning objectives of particular groups. Other studies have looked at metadiscoursal clues in lectures extending to multimodal features (Bernad-Mechó, 2018). Deroey (2017) compared the use of one type of discourse signpost – importance markers – in 160 authentic (un-adapted) lecture transcripts from four disciplinary groupings in a range of EAP coursebooks. She recommends that EAP practitioners use research-informed judgements about authentic language use to maximise the pedagogical value of listening material. However, the demands of measuring the performance of the 'listening phenomena' and its construct (the complex academic context in which the performance is enacted) in addition to appeals for the need for authenticity can have a deadening effect on an assessment writing team.

Graduateness: A tall order for a short course?

ESAP needs analysis is focused on the language and skills of a specific discipline. The purpose of research into graduateness is to articulate and measure the skills and attributes that students with a higher education in all disciplines can do better than the general population. Pre-sessional students are progressing to courses which develop graduateness skills as part of their disciplinary studies. These skills and attributes are often expressed in the grammar of learning objectives, but there is a debate about their precise nature, a debate given vitality by the need to align them across international systems as part of the Bologna process. Based on a review of the literature, Coetzee (2014) has developed a 64-item questionnaire with 8 core skills and attributes divided into 3 holistic domains of personal and intellectual development (see Figure 1). Similarly, Steur, Jansen and Hofman (2016) have proposed three domains: Scholarship, Reflective Thinking and Moral Citizenship; however, they wanted to describe a process of development such that a students' intellectual growth could be measured throughout their undergraduate and postgraduate education. Based on a reading of the literature, they suggest that reflection is the essential capability and that, as students are studying disciplines which draw on the component skills differently, progress in the other two may differ, but still represent a range and level of skills that demonstrate having had a higher education.

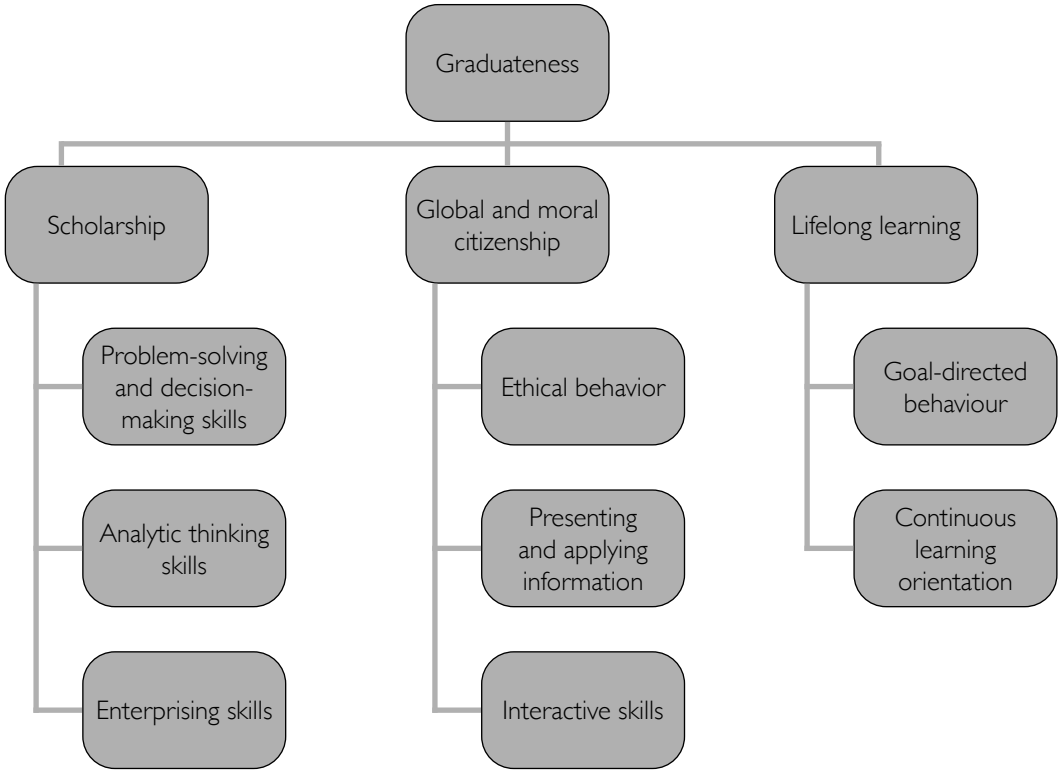

Figure 1 A classification of graduate attributes. Source: Coatzee (2014)

Kreber (2014) in a paper which is more philosophical, but retains a focus on actual student experience, talks of 'the need for students to develop the capacity for continuing learning in a world that is uncertain, having an inquiry-orientation and being capable of contributing effectively to civic life in a global context' (Kreber, 2014, citing Hughes and Barrie, 2010). She notes the sense of 'strangeness' that students feel when encountering the epistemological uncertainty of different disciplinary frameworks and practices, and argues that this can be dealt with authentically by producing one's own understandings through reasoning. Students become conscious that socially constructed assumptions can be challenged, that other cultures are valued and that reasoning can be applied to the common good. For Kreber, these dimensions of graduates rely on developing dispositions and qualities to actually apply or enact the practical reasoning that students are taught both in their studies and beyond. A pre-sessional may be able to start developing these skills and dispositions.

In the next section, an example is given in which a course leader fostered some of these skills, attributes and dispositions through the content of listening tasks.

Listening assessment in action: Commitments and compromises

The pre-sessional described here had a cohort of approximately 300 international

students, who had target courses across the whole university. The majority of the target courses required a student to demonstrate a language level of IELTS 6.0, with 5.5 in each skill. The pre-sessional had 5- and 10-week versions with a shared suite of assessments; marks are confirmed at an assessment board and communicated to the university's admissions team in the last week of the course. In common with many pre-sessionals, it was hoped to maximise value by having a curriculum direction-of-travel towards ESAP as far as practically possible.

The syllabus designers found delivering ESAP provision to be practical for reading and writing. Students used a common coursebook (e.g., Thaine, 2012) when developing reading and writing skills, transferred these skills directly to reading sources about their target discipline (e.g., Kotler and Armstrong, 2017, for Marketing students) and used their notes in summative and formative reading-into-writing assessments (using a range of discipline-appropriate genres, e.g., extended definitions, case studies, etc.). Teaching and assessing presentation skills followed a similar pattern of skills development and source use; however, developing ESAP listening sources into pedagogical and assessment tasks was more difficult given the range of target courses, the complexities of developing listening tasks, and the lack of off-the-shelf materials.

Recognizing graduateness as part of ESAP needs analysis allowed the test designers to have a set of listenings about topics outside, but complementary to, the students' target discipline and within the learning objectives of all target courses. So, for example, an assessment journey for a particular Marketing student has included writing a case study on branding, doing a presentation on that case study and then sitting a listening test which is shaped by a listening test construct and created around graduateness-related topics. Such an approach gives a principled and learner-centred solution to the practical problem of writing listening tests on an ESAP pre-sessional.

Language testing is a field where there is nothing as practical as a good theory. Glen Fulcher (2010) provides a detailed and practical guide to the theory of good language test writing and is essential reading for course leaders who have to defend a commitment to the quality of their assessment (as is his website http://languagetesting.info). An initial step for the course leader on this pre-sessional was to document the constructs and the test specifications (see Spaan, 2006, for an accessible explanation of how to write test specifications). Arriving at and documenting the listening test specifications was important for the test designers and for accountability to UKVI auditors. A test format of three ten-minute monologic lectures each played twice and assessed using multiple-choice items was thrashed out. Arguments for more authentic test items and tasks in listening testing were in-theory persuasive, but were set aside in a practicality-based compromise. The limitations of multiple-choice items are well established; but so are the remedies (see Bailey, 2016). Authentic listening test items for notetaking and summarizing had previously been used on the course, but were difficult to reliably standardize and mark.

The content of the lectures fell under each sub-section of Coatzee's (2014) classification of graduateness (Figure 1). For the sub-section of ethical behaviour, a lecture about the thought experiment known

as the 'trolley problem' (see Cathcart, 2013 or Edmonds, 2014) was written. For scholarship, a research focus was developed with a lecture on questionnaire design (drawing on Robson, 2007). For lifelong learning, a lecture containing a detailed review of a book on the limits of markets was used (Sandel, 2012). Each lecture had carrier content falling within graduateness topics which was used to test the listening construct in an academic context at B2 level.

Audio clips with these characteristics were difficult to find. They had to be created. Scripting gives the test writer control of what is being tested using the construct. Scripts were drafted and piloted to include target language, particular topics, and an appropriate speed and pace in line with the test specifications. Scripts can, of course, be based on transcriptions of authentic lectures and, therefore, include authentic features such a digression or corrections. They were recorded in quiet conditions and assembled with Audacity, a free audio-editing tool.

A listening test which used a single genre (lecture) and gameable test items (multiple choice) may have produced negative backwash. Backwash is the effect of a test on the teaching and learning that goes before it and can be both positive and negative (Alderson, Clapham and Wall, 1997). There were two ways to mitigate negative backwash in this case. First, the test worked in combination with other integrated skills tests; speaking and listening were also assessed by a seminar simulation. Second, the syllabus was managed in-house, so with a small group of teachers and a five or ten-week scheme of work, teachers were mindful to teach to the learning objectives and not 'game the test', i.e., break the rules.

The pre-sessional assessments needed to indicate whether the student had achieved the required level of English (CEFR, B2). In a pilot study, the test writers benchmarked the final mark of their test to that of a published IELTS one: 100 students sat both under exam conditions. The tests had different specifications for listening and different test items (IELTS includes short answers). However, they were able to see that the pilot group got lower marks on the IELTS test than the piloted one and so we were able to revise both the scripts and test items. Item analysis provided a powerful method of evaluating multiple-choice items (based on Fulcher, 2010). In addition to revising the test, the team also built their writing competence through these processes and readily concur with Brindley and Slatyer (2002), who found task difficulty to be remarkably complex due to the 'interactions between text, task and learner variables'. It also became clear that alignment to the CEFR is similarly difficult, as Harsch and Hartig (2015) found in research using human judges. However, the writers found item analysis and benchmarking to be important and practical in helping to ensure assessment quality.

Conclusion

Course leaders need to maximise the value of their pre-sessional provision to international students. They are high stakes for every international student and the institutions that educate them. Targeted ESAP provision seems to benefit students, but is a challenge to assess for pre-sessional teams. Using graduateness as a guiding system for scripting the content of listening assessments may offer a practical way to address the difficulty of testing all students

in a cohort with one version of a test without stretching the validity of the test beyond acceptable limits. It is not offered here as a replacement for rigor in test writing, but as a consideration within the compromises of language test construction.

References

Alderson, C., Clapham, C., & Wall, D. (1997). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Alexander, O., Argent, S., & Spencer, J. (2008). *EAP essentials: A teacher's guide to principles and practice* (2nd ed.). Reading: Garnet Education.

Bailey, K. (2016). Multiple-choice item format. In C. Coombe (Ed.), *Assessment and evaluation volume. TESOL Encyclopedia of ELT*. New York, NY: John Wiley & Sons.

Bernad-Mechó, E. (2018). *A multimodal approach to metadiscourse as an organizational tool in lectures* [Unpublished doctoral dissertation]. Castellón de la Plana: Universitat Jaume.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing, 19*(4), 369–394.

Bruce, I. (2011). *Theory and concepts of English for Academic Purposes*. Basingstoke: Palgrave MacMillan.

Cathcart, T. (2013). *The trolley problem: Or would you throw the fat guy off the bridge?* New York: Workman Publishing.

Charles, M., & Pecorari, D. (2016). *Introducing English for Academic Purposes*. London: Routledge.

Coetzee, M. (2014). Measuring student graduateness: Reliability and construct validity of the Graduate Skills and Attributes Scale. *Higher Education Research & Development*, *33*(5), 887–902.

de Chazal, E. (2014). *English for Academic Purposes*. Oxford: Oxford University Press.

Deroey, K. (2017). The representativeness of lecture listening coursebooks: Language, lecture authenticity, research-informedness. *Journal of English for Academic Purposes, 34*, 57–67.

Edmonds, D. (2014). *Would you kill the fat man? The trolley problem and what your answers tell us about right and wrong*. Princeton, New Jersey: Princeton University Press.

Field, J. (2011). Into the mind of the academic listener. *Journal of English for Academic Purposes, 10*, 102–112.

Field, J. (2019). *Rethinking the second language listening test: From theory to practice*. Sheffield: Equinox.

Fulcher, G. (2010). *Practical Language Testing*. London: Hodder Education.

Harsch, C., & Hartig, J. (2015). What are we aligning tests to when we report test alignment to the CEFR? *Language Assessment Quarterly*, *12*(4), 333–362.

Home Office (2019). *Tier 4 of the points based system: Guidance for sponsors*. London, HMSO.

Hughes, C., & Barrie, S. (2010). Influences on assessment of graduate attributes in higher education. *Assessment and Evaluation in Higher Education*, *35*(3), 325–334.

Hyland, K. (2016). General and specific EAP. In K. Hyland, & P. Shaw (Eds). *The Routledge Handbook of English for Academic Purposes (Routledge Handbooks in Applied Linguistics)* (pp. 17–29). London: Routledge.

Kotler, P., & Armstrong, G. (2017). *Principles of Marketing, Global Edition* (17th ed.). London: Pearson.

Kreber, C. (2014). Rationalising the nature of 'graduateness' through philosophical accounts of authenticity. *Teaching in Higher Education*. *19*(1), 90–100.

Lynch, T. (2011). Academic listening in the 21st century: Reviewing a decade of research. *Journal of English for Academic Purposes*, *10*(2), 79–88.

Manning, A. (2018). *Assessing EAP: Theory and practice in assessment literacy*. Reading: Garnet Education.

Robson, C. (2007). *How to do a research project: A guide for undergraduate students*. Oxford: John Wiley & Sons.

Rost, M. (2002). *Teaching and researching listening*. Harlow: Longman.

Sandel, M. (2012). *What money can't buy: The moral limits of markets*. London: Penguin.

Spaan, M. (2006) Test and item specifications development. *Language Assessment Quarterly: An International Journal*, *3*(1), 71–79.

Steur J., Jansen, E., & Hofman, A. (2016). Towards graduateness: Exploring academic intellectual development in university Master's students. *Educational Research and Evaluation, 22*(1–2), 6–22.

Thaine, C. (2012). *Cambridge Academic English: An integrated skills course for EAP (Intermediate)*. Cambridge: Cambridge University Press.

Vandergrift, L. (2004). Listening to learn or learning to listen? *Annual Review of Applied Linguistics*, *24*, 3–25.