University of Bedfordshire

Title: Testing academic literacy in reading and writing for university admissions

Name: Martine Holland

Testing academic literacy in reading and writing for university admissions

Martine Holland

Centre for Research in English Language Learning and Assessment (CRELLA)
University of Bedfordshire

A thesis submitted to the University of Bedfordshire, in fulfilment of the requirement for the degree of MA in Applied Linguistics by Research.

Degree awarded in July 2019.

Author's declaration

I, Martine Holland, declare that this thesis and the work presented in it are my own and have been generated by me as the result of my own original research.

I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University;
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- Where I have drawn on or cited the published work of others, this is always clearly attributed;
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- Where the thesis or any part of it is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- None of this work has been published before submission.

Abstract

Currently university entrance decisions are heavily reliant on further education qualifications and language proficiency tests, with little focus on academic literacy skills that are required to succeed at university. This thesis attempts to define what academic literacy skills are and to what extent they correlate with three measures of university success.

To answer these two research questions, I first investigated what academic literacy skills are through a survey of the literature, university study skills websites and existing academic literacy tests, and from these results drew up a checklist for academic literacy test validation. I then attempted to validate a new academic literacy test through a mixed methods study: first by calculating the correlations between performance in this test and university grades, self-assessment and tutor assessment, then through a case study approach to investigate these relationships in more detail.

My tentative findings are that, within the humanities and social sciences, the academic literacy test is likely to correlate strongly with university grades, both in the overall results and in two of the four marking criteria: coherence and cohesion, and engagement with sources, with some possibility of correlation in the argument criterion. The fourth criterion – academic language use – did not correlate, but this may be an effect of this particular participant sample rather than the test itself.

I also suggest two areas that may be difficult to elicit under timed exam conditions: eliciting appropriate source use when sources are provided, and eliciting synthesis of ideas across two or more given sources.

Acknowledgements

First, my heartfelt thanks to Professor Cyril Weir, my supervisor, who tragically passed away during my degree. His wealth of experience and knowledge, and his patience and good humour, will be hugely missed by all in the field of language testing and beyond.

Next, my massive and heartfelt gratitude to Dr Daniel Lam, who stepped in to supervise me when Cyril was no longer able to. I can never repay him for his effort, energy and insight, and for never despairing in the face of my relentlessly quantitative attitude towards qualitative research!

My thanks also to Dr Sathena Chan, who gave me invaluable advice on statistical analysis and a fresh perspective during my writing up.

I must also thank my colleagues at Cambridge Assessment English for the expertise they lent me during my research: Dr Nick Saville, Dr Kevin Cheung, Dr Gad Lim, Graeme Bridges and Andrew Kitney; and members of the University of Cambridge faculty for their opinions on the suitability of materials for their students: Dr David Chisnall and Dr Charlotte Lee.

Finally, huge thanks to my family, to my friends Jo, Sayuri and Helen, and my partner David, for their encouragement, perspective, and constant flow of tea and biscuits.

Table of Contents

List of tables

List of figures

1 Introduction

1.1 Background and rationale

Some universities have expressed a need for an entrance exam to identify the students that will perform most successfully at degree level. While subject knowledge is one possible predictor of success, this causes practical problems for universities: the majority of applicants may have achieved the highest grade in their secondary-level qualifications; international students' qualifications may not translate well to the local qualifications framework; there are subjects where secondary-level qualifications are rare, such as philosophy or computer science. It can also be argued that subject knowledge is not the only requirement for tertiary-level study. There is an additional area of knowledge – academic literacy – which can be broadly defined as 'reading, writing, listening, speaking, critical thinking, use of technology, and habits of mind that foster academic success' (ICAS, 2012, p. 2).

Advanced reading and writing skills have been highlighted as key (Sebolai, 2016; Flower in Patterson and Weidman, 2013a). The Intersegmental Committee of the Academic Senates (henceforth ICAS, 2002) suggested that a lack of analytical reading skills in particular is a significant contributor to a lack of academic success, and that the majority of students entering Higher Education are unprepared for commonly-assigned writing tasks involving analysing arguments and synthesizing information from varying sources.

In the UK, subject-specific tests (A-levels) are the primary factor in admissions with no overt focus on academic literacy skills. Some universities interview applicants, but this is neither widespread nor necessarily subjective, and is extremely resource-intensive. The UCAS personal statement (a free-text section of the UK university application form, completed by almost all undergraduate degree applicants) may not be objective nor necessarily representative of the applicant; the same can also be true for references.

The only other objective measure commonly used in considering applicants is the language proficiency test (hereafter referred to as LP tests) used in the admission of L2 speakers of English. Some of these, most notably IELTS and TOEFL, explicitly aim to assess students' preparedness for university-level study. Other LP tests such as Cambridge Assessment English's C1 Advanced have been recently revised to give more of an academic focus (Khalifa and Barker, 2015). However, research into the correlation between academic-English-focussed LP tests and academic performance has so far been inconclusive (IELTS: see Ushioda and Harsch, 2011; Weir, Chan and Nakatsuhara, 2013; and others discussed in section 2.2; TOEFL: see Harsch, Ushioda and Ladroue, 2017; Weigle, 2011 and Sawaki and Nissan, 2009). It should be noted that such LP tests are tools for deselection (Weir,1983), that is, for narrowing the admissions pool prior to the formal admissions process, and therefore not intended for use in predicting university success.

In response to this problem, universities and other stakeholders have begun to produce tests explicitly designed to test academic literacy. Although academic literacy tests (ALTs) have existed for a few decades in the UK (to the knowledge of this author, the earliest still in use is the University of Reading Test of English for Educational Purposes – TEEP – below), they have recently begun to gain in popularity. Currently there are two in common use: The New York State Teachers Exam (NYSTE) used in the selection of teachers in New York State, and the Academic and Quantitative Literacy (AQL) test for undergraduate university entrance in South Africa. Both of these tests have a reading component with multiple-choice questions and the NYSTE also contains a reading-into-writing component requiring the production of three texts from sources in approximately two hours. As these two tests are not currently available for UK admissions purposes, UK universities are devising their own ALTs. The University of Reading uses the TEEP in admissions selection, and Cambridge Assessment English is in the process of developing an integrated-skills test for postgraduate L2 admissions.

Academic literacy is still a loosely-defined construct in that each test mentioned in the above paragraph interprets that construct in a slightly different way (see section 4.1.3). The Cambridge Assessment English Academic Literacy Test (hence CAEALT) targets academic literacy with the following features (author's analysis, following discussion with the test's creators):

- Integrated testing of reading and writing to simulate authentic university assessment
- A clear thesis to be presented and supported from the reading material
- Authentic, ungraded source material
- Faculty-specific versions
- Longer output texts (c.800 words) approaching the length of authentic output texts

Finally, as ALTs are still a rare phenomenon, there are very few published analyses of such tests, and these analyses are slight. Additionally, while lists of subskills required for university study are available (see ICAS, 2012, or Rosenfeld, Leung and Oltman, 2001), they are not designed for test analysis and as such include subskills which are not appropriate for test conditions, such as 'experiment with new ideas' (ICAS, 2012, p.38) – it is not clear how this would precisely be defined, or how it would be manifest under test conditions – or 'read text material with sufficient care and comprehension to remember major ideas and answer written questions later when the text is no longer present' (Rosenfeld et al, 2001, p.80) – memory may be a controversial inclusion in an academic construct. Thus, the researcher seeking to produce or validate an ALT does not seem to have a clear list of skills available against which their work can be compared.

The aim of this dissertation is first to arrive at a construct definition of academic literacy through review of a variety of sources, and to establish the extent to which the construct underlying the CAEALT reflects this definition. The second aim of this thesis is to examine the relationship between CAEALT scores and academic success in terms of criterion-related validity, defined as 'the extent to which test scores correlate with a suitable external criterion of performance with established properties' (Weir, in Shaw and Weir, 2007, p.229). Criterion-related validity, in the form of concurrent and predictive validity, is usually targeted in university admissions tests where there is an interest in drawing a relationship between entry criteria and university success (Fyfe, Devine, and Emery, 2017).

1.2 Researcher’s background

I am an employee of an English as a Foreign Language international testing organisation, where I have been responsible for production and quality assurance of exam materials across all skills and now work on the development of new assessment, with a particular interest in the academic domain. Prior to this, I was an English as a Foreign Language / English for Academic Purposes teacher with a Diploma in Teaching English to Speakers of Other Languages (Delta), and a writing and speaking examiner for both the Academic and General English versions of the International English Language Testing System (IELTS).

2 Literature review

2.1 Exploring the definition of academic literacy

2.1.1 Scope of academic literacy: differing approaches

Academic literacy is broadly defined by ICAS as ‘reading, writing, listening, speaking, critical thinking, use of technology, and habits of mind that foster academic success’ (ICAS, 2012, p. 2). Within that broad-brush statement, Lea and Street (1998) identify three perspectives: a study-skills approach, a genre-based approach and a critical socially-situated approach. Any ALT should choose which of these three approaches is to be targeted, based on the literature and issues of practicality.

A *study-skills approach*: This perspective is based on a unified concept of academic literacy, which covers all faculties and academic areas: writing is seen as a technical skill with distinct ‘atomised’ elements which students are required to learn to progress; differences from the standard are problems to be ‘fixed’ (see Lea and Street, 1998). This approach is typified by Patterson and Weidman’s (2013b) statement that ‘there must be some degree of commonality that applies to all types of academic discourse which then allows one to perceive of this kind of discourse as typically academic’ (p.116).

Many LP exams with an academic slant, such as IELTS Academic, follow this approach, with texts and topics suitable for a candidate with no specialist knowledge, and marking criteria with no allowance made for discipline-specific variations in e.g. writing style –

the implication being that for university entry, such variations are not significant enough to impair reliability or to disadvantage candidates from particular fields. This is a practical approach for exam boards – who can produce a single version for each session; for candidates – who do not have to take multiple exams if they are considering several majors; and for receiving organisations – who can directly compare candidates' results without concerns about version equivalence.

A genre-based approach: This perspective acknowledges that there are differences in writing norms – format, rhetorical style, metadiscourse and lexis – between faculties, courses or even tutors. It is defined by Lea and Street (1998) as:

> 'how to write specific, course-based knowledge for a particular tutor or field of study. Problems lie with a lack of familiarity with the subject matter of a particular discipline and how to write that knowledge in that discipline' (Lea and Street, 1998, p.164).

In this view, academic writing is seen as a discourse practice: students are educated in the expectations of the university culture and in the particular expectations of writing within a particular discipline, and may be expected to be 'fluent' in the languages of multiple disciplines.

The idea of academic literacy appearing to be faculty-specific is well borne out in the literature, with Murray (2016) highlighting the work of Hyland, Swales, Nesi and Gardner, and Rex and McEachen as seminal.

Rex and McEachen (1999) list as areas that can vary between disciplines:

> ‘concepts and associated vocabulary… rhetorical structures, the patterns of action… characteristic ways of reaching consensus and expressing disagreement, or formulating arguments, or providing evidence, as well as characteristic genres for organizing thought and conversational action. (Rex and McEachen, 1999, p.69).

Gardner's 2011 analysis of the British Academic Written English (BAWE) corpus identified different ways in which the language of different faculties varies: the use of the first person, use of key phrases such as 'this essay' or 'in conclusion', key collocations. Gardner also believes that methods of constructing arguments can be seen through analysis of most commonly used words across disciplines: Philosophy and English deal in certainty ('absolutely, certainly'), Law and Sociology are more relative ('arguably, better, consequently'). Some fields such as Law allow appeals to authority, while texts in Philosophy are reliant on their internal logical structure.

A critical socially-situated approach: This approach is supplementary to the genre-based perspective in that it acknowledges variety in genre and faculty conventions, but stems from critical discourse analysis and cultural anthropology (see Lea and Street, 1998) and centres on the impact of authorial identity and power relations in student writing, and on ‘ideologically inscribed knowledge construction’ (Lillis, 2003, p.195). Lea and Street (1998) highlight two key issues that students face, namely that many requirements for writing on different courses and on what constituted acceptable knowledge are left implicit.

Lillis argues that this model should be considered alongside the previous two models, as it has:

> ‘helped to foreground many dimensions to student academic writing which had previously remained invisible or had been ignored; these include the impact of power relations on student writing, the centrality of identity in academic writing, academic writing as ideologically inscribed knowledge construction, the nature of generic academic, as well as disciplinary specific, writing practices.’ (Lillis, 2003, p.195).

Of the three perspectives of academic literacy considered here, the study-skills approach is the most commonly targeted by academic-focussed exams, primarily for reasons of practicality. With Nesi and Gardner (2012) identifying thirteen major genres and twenty-two different assignment types, a fully-realised genre-based approach would be entirely impractical for international, or even large-scale, exams. The critical socially-situated approach would prove even more impractical in terms of assessment: Weir (in an unpublished report for Cambridge Assessment English) suggests that ‘it is impossible to cater for this [perspective] in anything but a highly specific and limited way… as indicated by the case studies its advocates provide’ (n.d., p.12).

Weir (1983) argues that the study-skills approach is sufficient for the primary purpose of a LP test in university entry, that purpose being a deselection mechanism to ensure potential students can cope from a linguistic perspective. A compromise approach used by some tests targeting academic skills, such as that previously taken by the English Language Testing System (ELTS) and now proposed by the CAEALT, is to group subjects together by faculty: the latter offers papers tailored for Business and Administration, STEM and Humanities. It seems likely that this compromise provides a higher-level of context validity than those which do not relate to a specific faculty (see section 4.1). However, this distinction is only at the level of given texts and the question to be answered, rather than marking criteria, task type or examiner training / experience, and for this reason these tests can be considered as still following the study-skills model.

Given the use of the faculty-specific study-skills test type, the role of two other areas of literature must be considered in establishing the academic-literacy construct: English for Academic Purposes and the Reading-into-Writing construct.

2.1.2 English for Academic Purposes

English for Academic Purposes (EAP) can be broadly defined as 'those communication skills which are required for study purposes in formal education systems' (English Teaching Information Centre, in Jordan, 1997), a definition that can be directly compared with Sebolai's (2016) description of academic literacy as 'language competence that students need to possess in order to cope with the demands of academic study' given 'the technical nature of an academic linguistic sphere' (p. 46). While EAP and academic literacy can be seen as interchangeable terms which cover the same set of skills, in the literature review for this thesis the former tended to be applied to L2 speakers while the latter was equally applicable to English native speakers, implying variety of context, teaching techniques and expectations. Readers from the English as a Foreign Language field may find the inclusion of L1 speakers in this discussion unusual. However, Murray (2016) comments that a good level of academic literacy is 'something with which few if any undergraduate students, whether domestic or international, will enter university sufficiently equipped' (Murray, 2016, p.89). To emphasise this, this thesis will prefer the term 'academic literacy'.

EAP is sometimes divided into ESAP – English for Specific Academic Purposes (i.e. language for particular academic specialisms, such as medicine) and EGAP – English for General Academic Purposes (Blue, 1988). The latter category is further subdivided into study skills versus language features: study skills include areas such as, for speaking and writing, note-taking or giving presentations or, for reading and writing, appropriate referencing of sources or effective reading strategies. The term also covers non-skill-specific areas such as independent learning and revision strategies (University of Kent, 2017). The language features covered under EGAP were summarised by Jordan (1997) as 'a general academic English register, incorporating a formal, academic style, with proficiency in the language use' (p.5). Thus, a student successful in EAP (and, for the purposes of this thesis, academic literacy) would be familiar with the specific language necessary for their academic specialism, equipped with the study skills necessary to function in the university environment, and with a high level of proficiency in the academic English register.

In this thesis I will aim to define academic literacy using the approach of EGAP, and will primarily focus on the relevant study skills and appropriate academic English language use rather than subject-specific skills.

Following on from the conclusions in the previous section, literature from the EAP field reinforces the conclusion that an ALT should target the EGAP field i.e. that a study-skills test would be in line with EAP literature. The concept of *integrated reading-into-writing* is explored in more detail in the following section.

2.1.3 Academic English as integrated reading-into-writing

Many authors have aligned academic English and integrated reading-into-writing (see Cumming et al, 2006; Knoch and Sitajalabhorn, 2013; Plakans, 2010). ICAS (2002) go so far as to say that in the academic context 'no one disputes the connection between reading and writing[...]. Students […] should articulate a clear thesis and should identify, evaluate, and use evidence to support or challenge that thesis'. (ICAS, 2002, p.15). Cumming (2013) notes that

> 'the integration of content from source material…is what writing for academic purposes involves. Students at schools, colleges, or universities are mainly asked to write in order to display their knowledge of ideas and information from reading… as well as their abilities to analyse and communicate this material purposefully and coherently' (Cumming, 2013, p.3)

Therefore, the argument for the use of integrated skills in tests is primarily based on increased contextual authenticity. Chan (2013) argues that reading-into-writing tests 'better represent the performance conditions of real-life academic tasks' and that the 'processes writers employ when they write from sources […] are important for academic writing, but these processes seem to have received little or no attention in most current writing tests' (pp. 32-33).

It also appears that this task type elicits more authentic communicative functions: Cumming et al (2006) found that candidates' texts from reading-into-writing tasks were more reliant on other sources of information and less on exhortation, as well as displaying a wider lexical range and longer clauses.

The above has argued that, when assessing reading and writing skills, integrated reading-into-writing testing is more valid than testing each separately. There is also an argument that academic skills are best represented by reading and writing rather than listening or speaking, as the former are more representative of the academic sphere; that is, academic reading and writing are more distinct from general language proficiency than academic speaking or listening. Sebolai (2016) notes that while the dominant modes in general English are speaking and listening, academic English is often based much more on the written language and therefore on the corresponding language functions – classifying, comparing, contrasting and inferencing – as well as placing a greater emphasis on cohesion and coherence, making the corresponding lexis and syntax necessary for academic work. Flower (in Patterson and Weidman, 2013a, p.3) agrees in that 'integrating information from sources with one's own knowledge and interpreting one's reading/adapting one's writing for a purpose' are 'practices that seem to be critical features of academic language'. However, given that the literature in this area is slight, there is not yet enough evidence in this area to exclude speaking and listening from admissions testing.

Chan (2013), in her research comparing two reading-into-writing exam tasks and real-life reading-into-writing tasks on a number of parameters, concluded not only that (as discussed above) a reading-into-writing test was a reasonably accurate reflection of real-life university activity, but also that the reading-into-writing test cognitive construct is not simply an amalgamation of the two separate skills, but a construct in its own right. In particular, she introduced another stage in the integrated cognitive construct: meaning and discourse construction, which included the substages connecting and generating, selecting relevant ideas and careful global reading. This is in line with Plakan's (2010) research into learners' task representation of integrated and independent tasks, which found that the integrated tasks required monitoring to (1) ensure effective synthesis and (2) avoid plagiarism, as should be the case with academic writing.

A few concerns have been expressed with the reading-into-writing format:

- Khalifa and Weir (2009) highlight the two possibilities of extensive lifting of input material and a lack of development of ideas presented, although this can be mitigated by giving the candidates restrictions on lifting 'as in real life rules concerned with plagiarism' and requiring 'input language transformation' from the candidate. (p.91).
- Plakans (2010) notes that candidates may not be familiar with the synthesis required for this task type, and that even clear directions as to appropriate source use may not be sufficient to elicit this. This is a particular issue for an admissions test which does not follow a specified course.

- Chan, Wu and Weir (2014) note that a major gap between real-life and academic-writing tests is the former's use of multiple extended secondary sources. This is impractical in exam conditions, and therefore represents an unavoidable gap in the context validity of this exam format.

In summary, integrated reading-into-writing is considered to be more contextually valid and to elicit more authentic language. It also has a distinct cognitive construct, which includes meaning and discourse construction. For these reasons, my study will focus on integrated reading-into-writing in the academic context.

***

The research presented in this section suggests that an ALT should target 1) the study-skills approach, as the most practical approach; 2) academic literacy as the specific functions, syntax and vocabulary necessary for the academic environment; 3) the reading-into-writing response format, as both the most valid form of academic skills assessment and the most representative of assessment in the academic environment.

2.2 Measuring the success of admissions tests

In the discussion above, I have explored the literature surrounding the construct and definition of academic literacy, and shortly will relate this literature to the admissions test context. Another key area of literature refers to whether admissions tests, of which the CAEALT is an example, can successfully predict performance at university – that is, how effective admissions tests can be.

As discussed in section 1.1, those researching the effectiveness of admissions tests are often concerned with predictive and concurrent validity: that is, with correlation of test scores with another measure of what is believed to be / has a proven track record of being the same ability (Milanovic and Weir, 2009). Shaw and Weir (2007) give the definition of concurrent validity as 'comparing scores from a given test with some other measure of the same ability of the candidates taken at the same time as the test' (p.229), and predictive validity as comparing 'test scores with a measure for the same candidates taken some time after the test' (p.229). That is, the primary difference is one of timing: the former compares two measures taken at approximately the same time, while the latter compares the test scores with another measure taken at a point in the future.

In choosing a predictive-validity study over a concurrent-validity study, or vice versa, there are several issues to consider. Where a study proposes a new tool to substitute for one that already exists, both of which test the same or a similar construct, concurrent validity is most commonly chosen (Fyfe, Devine, and Emery, 2017). Where such a widely-accepted tool is not available, predictive validity is preferred. In the case of university admissions, no widely-accepted tool exists, and so 'establishing good predictive validity is often seen as the holy grail of admissions tests' (ibid, p.144).

As well as there being no clear tool for admissions testing, there is no widely-accepted measure of university success. Measures that have been used by various researchers with good levels of success are university GPA (Yen and Kuzma, 2009; Humphreys et al, 2012); coursework grades (Ushioda and Harsch, 2011): academic tasks (Weir, Chan and Nakatsuhara, 2013); student self-assessment (Kerstjens and Nery, 2000) and tutor assessment (Cotton and Conrow, 1998; Ingram and Bayliss, 2007). Cotton and Conrow (1998) list other possible measures including course or module pass or fail and the amount of work successfully completed. In particular, Pollitt (1988) considers student self-assessment to be an important measure, as it 'constitutes potentially the broadest and most valid of all proficiency assessments if it can be made sufficiently reliable' (Pollitt, 1988, p. 63). Pollitt does not go into greater detail as to why, but it can be expected that a student may have the most accurate idea of their progress, assuming that they have progressed far enough on the course to have received regular feedback from tutors.

Finally, predictive validity studies can be more difficult to carry out than concurrent validity studies (See Fyfe, Devine, and Emery, 2017, p. 145). For further discussion of the issues, please see section 2.2.2.

2.2.1 Language proficiency tests

There is a significant body of literature on the predictive and concurrent validity of LP tests. This literature is inconclusive as to the predictive power of these tests. To take IELTS as an example typical of the field, some studies have found a positive correlation with different measures of university performance – Feast (2002) finding a correlation of 0.39 between IELTS and GPA; Ushioda and Harsch (2011) finding 0.38 between IELTS and coursework grades; Weir, Chan and Nakatsuhara (2013) finding 0.41 between IELTS and four set academic tasks – while others have found no clear correlation (Ingram and Bayliss, 2007; Dooey, 1999).

Correlations between individual skills fare similarly: Humphreys et al (2012), in comparing an IELTS exam taken at the end of the first semester with the first semester's GPA, found a correlation for listening and reading (0.34, 0.34), but non-significant results for writing and speaking. Ushioda and Harsch (2011) found a correlation for reading, writing and listening (0.50, 0.47, 0.38), but not for speaking (0.26). There is also evidence to suggest that these ambiguous results may to some extent be dependent on the level of proficiency: Cotton and Conrow (1998) found that IELTS performed as a better predictor of course marks at lower levels than higher.

Murray (2016) notes that these low correlations do not necessarily suggest that LP tests should not be used, as 'alternative practicable means… are hard to discern' (p. 107) and that they do a 'reasonably good job' (p.107) of suiting the purposes of universitie s and students. Cho and Bridgeman (2012), in their comparison of the TOEFL iBT to academic performance which found a small correlation between the two measures, emphasized that, due to the issues discussed in section 2.2.2 below, 'even small correlations or seemingly trivial amounts of variance explained may be an indication of a meaningful relationship between two variables' (ibid, p. 439). Another argument for taking low correlations as meaningful is that the use of LP tests is distinctly different from other university admissions tests in that there is no expectation that such tests will be used other than as a tool for deselection. That is, universities specify the minimum LP requirements to be able to deal with the linguistic demands of their course; a higher

IELTS, TOEFL or similar score will not necessarily correlate with a higher final university mark.

Due to the lack of correlation between IELTS scores and university performance described above, Ingram and Bayliss' (2007) study focussed on the ability of IELTS scores to predict the language behaviour of students at university level and how well these language behaviours were able to cope with the tasks set at university. Their research suggests that the LP variable is a contributing factor to academic performance, particularly in more linguistically demanding subjects.

2.2.2 Predictive and concurrent validity: key issues

In the previous section I introduced predictive and concurrent validity studies and some of the outcome measures used in such studies. However, there are two key issues that must be considered in making such design choices and in interpreting the results of studies.

Range restriction

According to McManus and Dewberry (2013), a key problem faced by those estimating predictive validity in gatekeeping exams is that, 'while selection takes place in the entire pool of candidates or applicants, validation of the predictor measures can only take place in those who have entered [the accepting institution]' (McManus and Dewberry, 2013, p.4). This restriction in range has the effect that the correlation calculated between the target attribute and the accepted student results is significantly lower than it would be if results were available for the entire pool of applicants, or even for the entire population of the country. Two other factors that reduce the possible correlation is right-censorship –the highest achieving students cannot achieve higher than the maximum mark available for the test – and that consistently high cohort performance may lead to little variation in the admitted cohort (See Bridgeman, Cho and DiPietro, 2016).

Chernyshenko and Ones (1999) found that the use of appropriate range restriction had a dramatic effect on their validity measures, which more than doubled and thus demonstrated the exam under consideration was in fact a good measure of performance.

Confounds on the outcome variable

Some authors (Bridgeman, Cho and DiPietro, 2016; Ingram and Beyliss, 2007) account for the lack of correlation in predictive validity studies by the presence of many different confounding factors, saying that 'most predictive studies based on language tests… can be criticised on the grounds that it is impossible to account for all the variables' (Ingram and Bayliss, 2007, p.5). The issue of confounding factors becomes greater the more distant in time the two measures are from each other, with consequences for research design: for example, while end-of-course grades may be seen as the obvious choice as a measure of academic ability, because admissions tutors and students alike are aiming for successful completion of the course, an outcome three years distant from selection allows for many possible confounding variables.

The outcome variable selected as a criterion can be confounded in various ways, for example:

1. Different departments, colleges or tutors may offer different levels of teaching quality, or different levels of marking reliability: combining such variety into one outcome score was highlighted by Bridgeman, Cho and DiPietro (2016) as a common reason for lower or even negative correlations in previous studies.

2. Continuing from the above, Murray (2016), an advocate of the genre-based approach to academic literacy, writes that 'high-profile gatekeeping tests… focus on generic EAP… this fails to take account of the particularity of literacy practices within specific disciplines' and that 'performance [at university] is largely dependent on students' conversancy in those practices pertinent to their particular disciplines and with which…we cannot assume or expect students to come equipped to university' (Murray, 2016, pp. 106-107).

3. Cotton and Conrow note that variety in measurement, even for such a standard measure as GPA, means that correlations between two scores can be difficult to interpret, and therefore that 'it is therefore important to use more than one measure of academic achievement in predictive validity studies' (Cotton and Conrow, 1998, p.76).

4. A poorly-performing student may be offered further support. Fyfe, Devine, and Emery (2017) note that such support is particularly common for written

communication skills.

5. Cotton and Conrow (1998) highlight the importance of student motivation as a key variable to be considered in the analysis of results.

6. Socioeconomic status is highly correlated with both exam results and university performance and therefore, when ignored, can dramatically overinflate the predictive power of exams (Atkinson and Geiser, 2009).

In summary, there are many issues to consider when calculating predictive or concurrent validity for university-entrance aptitude tests. Correlations are difficult to measure (due to confounding variables) and also significantly lower than in other fields (due to range restriction). Multiple measures of ability are helpful in attempting to counteract such issues. The methodology in this study has aimed to reduce confounding variables where possible, by accepting participants from the Humanities and Social Sciences only, by recruiting as many participants as possible from one university, by using the same marker for all scripts, by using a marker versed in writing for Humanities and Social Sciences, and by taking multiple measures of academic success.

In this literature review I have addressed the literature in two fields: first, that of the construct and definition of academic literacy, and second, that of the criterion-related validity of admissions tests. There appears to be a need for a taxonomy of the academic construct, at a specific and granular level, designed with test makers in mind. I have also discussed issues surrounding the problems that need to be addressed in producing and validating an admissions test.

Thus, this thesis aims to answer two research questions:

RQ1a: What is a suitable construct of academic literacy, in the context of the humanities and social sciences, to be targeted in large-scale undergraduate admissions testing?

RQ1b: To what extent is the Cambridge English Academic Literacy Test (CAEALT) representative of this construct?

RQ2: What is the relationship between performance in the CAEALT and academic success?

3 Methodology

3.1 RQ1: methodology

The first part of RQ1 seeks to define the construct for academic literacy. Figure 1 shows the process used.

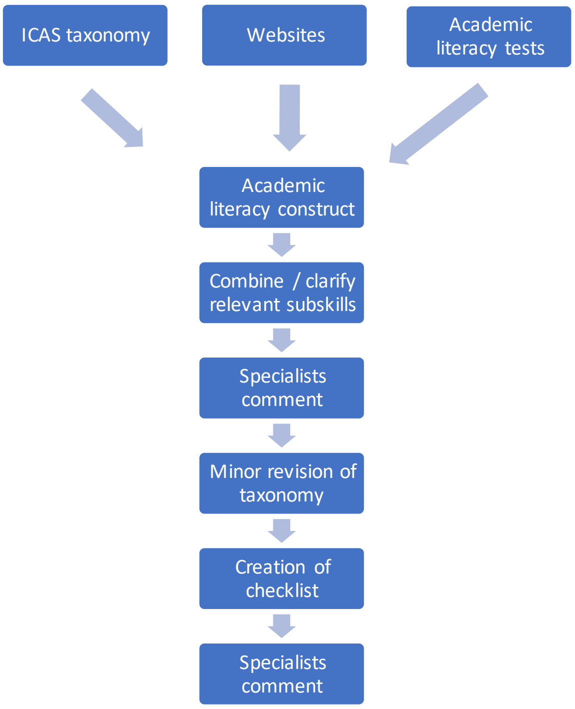

Figure 1: Process of producing the taxonomy and checklist

3.1.1 Surveying materials

The starting point for the taxonomy was the ICAS academic literacy statement of competencies (ICAS, 2002). This statement consists of first a discussion of the competencies required for university, and then a list of subskills, divided into categories such as reading competencies, technology competencies and so on, arrived at through a survey of academic faculty members in California.

From this document, competencies were selected under the following headings: academic literacy and critical thinking, making the reading/writing connection, reading competencies, and writing competencies. As the goal of this taxonomy is analysis of ALTs, competencies that, in my opinion, could not be easily tested under controlled examination conditions were removed from this list. These include, for example, the competencies 'suspend information while searching for answers to self-generated questions' (p.39) 'have strategies for reading convoluted sentences' (p.40) or 'use the library catalog[ue] and the Internet to locate relevant sources' (p.41).

This list of competencies was compared against sources in three key areas, making additions as necessary:

A. **research literature** in the fields of academic literacy, university entrance examinations and LP testing

B. study-skills pages on three **university websites**, which explicate the skills expected of their students

C. the constructs of two **academic literacy tests**: the Reading TEEP and the New York State Teachers' Examination (NYSTE) Academic Literacy Test

These three key areas lay out three common but different perspectives on academic literacy: the research into relevant areas, which provides a theoretical and evidence-based point of view; the skills required of students from the point of view of universities, whose webpages are likely to be written in response to their particular student need, thus providing a broader overview of skills required by students across academic institutions; and those skills as interpreted by test makers, which, due to the nature of test-making, are likely to be more concrete than the other sources. These three sources together mean that the taxonomy created is likely to be evidence- and research-based, broad, and suitable for test creation. There are of course areas of overlap in that the research will be consulted by (and in some cases commissioned by) test makers, and that both the ICAS statement of competencies and the websites elicited the views of teaching faculty.

I surveyed fifteen university websites in total: they were chosen by an internet search under the following key terms: study skills, reading skills, and how to write an essay / dissertation / research proposal (as a result of these search terms, only websites in English were considered). The three chosen to contribute to the taxonomy were the Open University, the University of Kent and Dartmouth College; these had a high level of granularity (discussing reading and writing skills in detail) but also carried some measure of authority (the sites chosen had more than one mention in the generated internet search, and the Open University website was referenced by several other websites).

The two ALT constructs (Reading TEEP, NYSTE) were chosen for inclusion in the taxonomy based on having a written productive component (ruling out tests such as the South African Academic Literacy Test), and explicitly aiming to test academic literacy as opposed to general LP (ruling out tests which self-identify as English-language tests, such as the Canadian Academic English Language Test, IELTS or TOEFL). This selection also includes a test from each of the two main categories of ALTs: those specifically designed for L2 university entrance, and those intended to test academic literacy across L1 and L2.

3.1.2 Scoring criteria

To manifest the relative importance of each subskill, each was given a score out of 5 for each ALT and the ICAS taxonomy (a total score of 15 for each subskill) and a score of 4 for each of the websites (a total score of 12). This difference in total possible scores between the tests (10) and websites (12) was partly an effect of the different granularity possible in the websites as opposed to the tests, but also to ensure that the tests and websites were approximately equally represented in the final taxonomy, despite the difference in number of tests/websites consulted.

For the tests and ICAS taxonomy, a score of 5 indicates that the subskill is clearly tested in the test construct or explicitly stated as part of the construct in the exam board's accompanying literature; 2.5 indicates that it is tested to some extent or implicit in the literature; 0 indicates that it is not tested and not mentioned in the literature. For university websites, 4 indicates that the subskill was given an entire page on the website; 3 indicates it was a key idea on at least one page; 2, that it was mentioned on at least one page; 1, that it was implicit; and 0, that there was no mention of the subskill on the website. For example, the subskill 'identify authorial attitude' is listed in the ICAS taxonomy (giving a score of 5); it is required by the Reading TEEP for accurate synthesis of the sources but not mentioned in the accompanying literature, giving a score of 2.5; it is explicitly mentioned in the accompanying literature for the NYSTE (New York State Education Department, 2014), giving a score of 5. However, this subskill is only implicit in two of the three websites (giving a score of 1 for each), and mentioned in passing in the third (Open University), giving a score of 2. Therefore, the total score for this subskill was 16.5.

ICAS taxonomy	5
Reading TEEP	2.5
NYSTE	5
Website: University of Kent	1
Website: Dartmouth College	1
Website: Open University	2
Total	16.5

Table 1: Scores for the subskill 'identify authorial attitude'

3.1.3 Expert judgement and subskill selection

The overall taxonomy produced through the survey of materials (ICAS, websites, ALTs) was then sent to two specialists for comment: one from the field of writing and reading-into-writing testing, and the other from the field of university-admissions testing. I invited the specialists to comment in all areas, but particularly regarding the completeness of the construct, based on their expert knowledge.

At this point I and these specialists reached agreement on whether all the subskills on the taxonomy were necessary in an ALT or whether a smaller subset was more appropriate. Section 4.1.5 describes the decision-making process.

3.1.4 CAEALT analysis

To answer the second part of RQ1, the extent to which the CAEALT is representative of this academic literacy construct was analysed against the checklist. Each subskill was assigned one of three categories: Y – definitely required to complete the task, P – possibly required to complete the task, and N – not required. Each category was assigned a score to allow the calculation of a total (Y=5, P=2.5, N=0); this total represents the extent to which the CAEALT represents the construct drawn up in this thesis.

This analysis was repeated by a writing materials production specialist, resulting in no changes to the assigned scores.

3.2 RQ2: methodology

This study aimed to explore the relationship between performance in the CAEALT and academic success, using a mixed-methods sequential explanatory design, where I first collected quantitative data on CAEALT performance and measures of academic success, as well as qualitative data in the form of candidate feedback on the test and case studies analysing a selection of candidate scripts, to further explore the relationship between the test performance and academic performance and gain insights into the academic literacy construct assessed by the CAEALT.

3.2.1 Research instruments

The following instruments were used for the study: The CAEALT and resulting candidate scripts; self-assessments, tutor assessments and self-reported grades (as measures of academic success); feedback on similarities between the candidates' everyday university life and the CAEALT.

The CAEALT has three variants: Business and Administration, STEM, and Humanities. It consists of sources and an essay question. The sources are two longer texts (1000+ words) and tables or infographics. These texts are suitable for reference in academic writing, and can include journal articles and textbooks. Bibliographies for each source are also presented to allow secondary referencing. The candidate has 2.5 hours to produce a text of approximately 800 words, which must engage with the sources. The markscheme rewards critical engagement with the sources and accurate referencing, as well as strength of argument, appropriate coherence and cohesion and accurate and varied academic language use. Unfortunately, due to reasons of confidentiality, the markscheme cannot be included in this thesis.

RQ2 is a concurrent validity study: due to practical constraints (see section 3.2.4),
it was not possible to take a consistent measure of future performance to be compared with the CAEALT. As a concurrent validity study, it was necessary to choose measures of university study ability that corresponded as closely as possible to the construct targeted by the CAEALT, i.e. academic literacy separate from subject knowledge or other confounding variables; it is also important to use more than one measure of academic

achievement in criterion-related studies (see section 2.2.2). Thus, three proxies were chosen as measures of university success:

- University grades from the end of the candidate’s previous year: Previous studies have used this measure with success (Yen and Kuzma, 2009; Humphreys et al, 2012) and while this does introduce confounding variables, the high-stakes nature of this measure suggests it can be considered the most reliable of the three included here. End-of-year scores were self-reported (appendix 8.6) and an official transcript of all modules was also requested when available to verify these results.

- Participant self-assessment (appendix 8.4.5): a measure used by Kerstjens and Nery (2000), among others, this measure is particularly valued by Pollitt (1988) (see section 2.2). While questions can be raised about this measure’s reliability, students who have received regular feedback on their progress are likely to have a greater understanding of their ability. This questionnaire consisted of the subskills identified in the RQ1 checklist, slightly reworded to make them more accessible to students. Candidates were asked to rate themselves against these subskills on a 5-point Likert scale. Both this questionnaire and the tutor assessment below were checked beforehand for comprehensibility by a Director of Studies (a member of academic staff responsible for the academic welfare of all students in their subject at their college).

- Tutor assessment (appendix 8.4.6): a measure of achievement used by Cotton and Conrow (1998) and Ingram and Bayliss (2007). Both self- and tutor assessments were used to include different perspectives on the same subskill, as well as to examine any differences between ratings. As with the self-assessment, this measure consisted of a questionnaire based on the checklist of subskills, reported against on a 5-point Likert scale. The subskills targeted were those which could be reported against by tutors: subskills 2.1.2 (reading strategies) and 6.2.1 (revision techniques) were considered unsuitable for tutors to report against.

The final instrument used was a form for candidates to identify ways in which the CAEALT was similar to and different from their everyday university studies. This information may provide an insight into correlations between the other measures by identifying:

- Areas of the construct not covered by the test
- Key differences of context between the test and university essay assessments
- Any areas where the subject studied by the participant may skew the result. For example, a participant may comment that the task type in the CAEALT is not a standard form of assessment on his or her course.

In summary, RQ2 intended to identify and illuminate the strength of the relationship between academic success and performance on the CAEALT by taking multiple measures of academic success, each of which provides a slightly different perspective on the candidates' performance, and by gaining an understanding of the candidates' everyday university study experiences.

3.2.2 Participants

The primary restriction on participant demographic was that of subject: the examiner who marked the tests had a Humanities and Social Sciences background, and the CAEALT has a Humanities and Social Sciences variant. This limitation would also help to mitigate possible issues of genre-based differences in marking standards and ensure reliable marking. The secondary consideration in selecting participants was to minimise confounding variables where possible (see section 2.2.2). For this reason, the criteria for participation began narrow and gradually broadened as difficulties in recruitment became apparent.

Potential participants were contacted via department email circulars. When potential participants made contact, if they met the criteria for participation, they were informed of the purpose of the research and the time and data required of them. A £20 gift voucher was initially offered as an incentive to participation; this was increased to £30 after the first two months to increase take-up.

Students from the University of Cambridge were initially targeted for two reasons. The first and primary reason is that it is an instance of a highly-prestigious university: as discussed in the introduction, academic literacy tests may be of substantial use to universities where a large proportion of applicants have received the highest possible grade in relevant achievement tests (such as A-levels), and which thus require an additional discriminating factor. The second reason was practical: I knew several

Directors of Studies at this university and so thought there was a higher probability I would be able to get the required number of participants.

As such a choice of top-tier institution risks limiting the applicability of this thesis, alternative student cohorts were considered at this stage such as including students from universities with lower entrance requirements, or working exclusively with these universities. However, I rejected these options: as well as the reasons listed in the previous paragraph, it was important for methodological reasons to minimise confounding variables (such as marking criteria, band scores or disciplinary practices) by limiting the number of institutions and courses.

Undergraduate students were initially preferred to postgraduates: while the test was designed for University of Cambridge postgraduates, I believed it would also provide insights for an undergraduate population. This hypothesis was seconded by the designer of the CAEALT, and the suitability of the test for undergraduate students was checked by a Director of Studies for Modern Languages at the University of Cambridge.

No preference was given for L1 or L2 participants. The reasons for this are partly practical: the number of volunteer participants was highly limited, particularly the number of L2 volunteers. However, there are also theoretical grounds for this decision: while the concept of academic literacy is often discussed within the L2 testing context, many authors in the field (ICAS, 2002; Lillis, 2003; Nesi and Gardner, 2006; among others, as well as the NYSTE) view academic literacy as a skill to be acquired by L1 and L2 speakers alike (see also section 2.1.2). Table 2 details the LP of the participants.

Language proficiency	Number
L1 (born in English-speaking countries)	12
L1 (self-identifying as English bilingual, born in non-English-speaking countries)	2
L2 with previous LP scores	3 (IELTS 8, IELTS 7.5, Cambridge Proficiency B)
L2 without previous LP scores	1

Table 2: First language of participants

3.2.3 Data collection procedures

	Quantitative data	Qualitative data
Before test administration	Demographic data	
	Participant self-assessment	Additional comments from participant self-assessment
	End of first year results	
Test administration	CAEALT results	
After test administration		Test-taking experience questionnaire
	Tutor feedback	Additional comments from tutor feedback

Table 3: data collected

Data was collected as follows (see Table 3 for a visual representation of this process).

- **Before test administration**

Permissions were requested from the participant before any data was collected.

Before test administration, candidates provided demographic data, a self-assessment and end-of-year results. Requested demographic data consisted of gender, nationality, course and faculty, college (within the University of Cambridge) / university, previous educational attainment, previous LP testing experiences, previous LP test scores, reading and writing LP scores (if available). The participant was also asked to nominate an appropriate tutor to be contacted for the relevant data collection.

- **ALT administration**

Test administration took place under strict exam conditions. Exam scripts were marked by one examiner, with six scripts selected at random for re-marking at the end of the marking process (no marks were changed as a result of this). To help ensure consistency of marking, the examiner used familiarisation materials produced by Cambridge Assessment English prior to marking.

- **After test administration**

Finally, candidates completed the form identifying similarities and differences between the CAEALT and their everyday university studies.

Following the test day, tutor assessments were requested.

3.2.4 Issues with participant recruitment

There were significant problems in participant recruitment: The hope was that thirty participants would be available and collection would take three months; in the event, eighteen participants were recruited over seven months. Data collection took place in three main stages: firstly, Humanities undergraduates at the University of Cambridge were targeted, beginning their second year. By choosing students at the beginning of their second year, I hoped to access end-of-first-year scores, allowing the calculation of correlations between CAEALT score and course grades at two points in time. These requirements were quickly broadened to both Humanities and Social Sciences, in either their second or third year. Secondly, other universities were approached, gaining the study two further participants. Finally, current and recent postgraduates were included, from a range of universities (see Table 4).

	University of Cambridge	Other UK universities	Overseas universities
Undergraduate students	9	1 x Anglia Ruskin 1 x Liverpool	0
Postgraduate students	0	1 X Edinburgh 3 x Lancaster 1 x Nottingham	1 x Melbourne 1 x Corfu

Table 4: Participants' universities

This study was initially intended to be a predictive-validity study, measuring candidates' academic literacy at two points: the beginning and end of the academic year. Unfortunately, participant recruitment took significantly longer than planned for, with many participants recruited in the second half of the academic year, meaning that a two-point measure was no longer applicable. Also, some of the postgraduate candidates had recently finished their degree and so no second point in time was available. Therefore, all results are based on the most recent university grades available at the beginning of the academic year or, for current postgraduate students, modules completed to date.

The inclusion of tutor feedback caused some issues: First, the necessity of broadening the participant pool to include postgraduate students meant that there were some who had recently finished their courses and were no longer in touch with their tutors, and some who had studied in an extremely large cohort with highly-limited tutor contact, meaning that tutor feedback would not be meaningful. Secondly, some students / tutors had confidentiality concerns and preferred not to give tutor names / feedback. When data collection was closed, assessments from five tutors had been returned.

The final participant demographic for the sample is given in Table 5.

	Undergraduate		Postgraduate	
	2nd year	3rd year		
Stage of education	6	5	7	
	M	**F**	**M**	**F**
Gender	3	8	4	3
Age				
19	2			
20	6			
21	1			
22	1			
23-30	1		1	
30-35			1	
36-40			4	
41-45			1	
First language				
English	6		6	
Punjabi	1			
Hungarian	2			
Spanish	1			
Polish	1			
Greek			1	
Faculty				
Humanities	3		2	
Education	3		1	
Social science	4			
Philosophy	1			
Linguistics			4	

Table 5: Demographic information for the sample

3.2.5 Data analysis

3.2.5.1 Quantitative data

Prior to analysis, self- and tutor assessment subskills were assigned to each of the four CAEALT marking criteria (argument, coherence and cohesion, academic language use, engagement with sources) by the researcher to allow analysis both overall and at a more granular level. These assignments were then checked by a writing assessment expert.

Data was analysed using the SSPS program. The mean and standard deviation were calculated, as well as correlations between each measure of academic performance and the CAEALT scores as listed below. Unreliability in the predictor scores was not corrected for, as this is inappropriate in measures used for university selection (See Chernyshenko and Ones, 1999).

Correlations were calculated using Kendall's Tau as the measure most appropriate for establishing a relationship between two measures when the assumptions necessary for Pearson (a normal distribution, a linear relationship between variables) or Spearman (a monotonic relationship between the two variables) do not hold. As is conventional, a result was considered significant with a p-value of smaller than 0.05. Effect size was calculated using Cohen's d, as the group size and standard deviation of each group were similar.

Measures thus calculated were:

- CAEALT marks and university grades: overall; by each marking category

- CAEALT marks and self-assessment: overall; by each marking category; by each individual subskill
- CAEALT marks and tutor assessment: overall; by each marking category; by each individual subskill
- Self-assessment and university grades: overall; by each marking category; by each individual subskill
- Tutor assessment and university grades: overall; by each marking category; by each individual subskill
- Self-assessment and tutor assessment: overall; by each marking category; by each individual subskill

The US Department of Labor, Employment Training and Administration (See Fyfe, Devine, and Emery, 2017, pp 177-178) give the correlations in table 18 as suitable for use in predictive validity studies.

Validity Coefficient	**Interpretation**
Above 0.35	very beneficial
0.21 to 0.35	likely to be useful
0.11 to 0.20	depends on circumstances
Below 0.11	unlikely to be useful

Table 18: Guidelines for interpreting correlation coefficients in predictive validity studies (Fyfe, Devine, and Emery, 2017, pp 177-178)

3.2.5.2 Qualitative data

To provide further supporting evidence for claims drawn from the quantitative data, two qualitative analyses were carried out. First, additional comments from the self- and tutor assessments and the test-taking experience questionnaire were analysed for common themes. Second, the script, candidate and the tutor for four participants and the test-taking experience questionnaire were analysed in depth to produce case studies, with the aim of gaining further insights into these participants' academic literacy and identifying any commonalities. The highest and lowest scoring participants in the sample were selected, and the analysis examined the strengths and weaknesses indicated in their CAEALT scripts and in the self-/tutor assessment.

***

From this point on, as the results of RQ2 will shed further light on RQ1, I will look at the results and discussion together for RQ1 before moving on to the results and discussion of RQ2. I will then revisit RQ1 in an overall discussion section (section 6), taking the concepts from RQ2 into account.

4 RQ1: Results and discussion

4.1 RQ1a: results

RQ1a: What is a suitable construct of academic literacy, in the context of the humanities and social sciences, to be targeted in large-scale undergraduate admissions testing?

RQ1a involves consulting three different sources of information on the scope of academic literacy: the literature, university websites and ALTs. This section will review the findings from each in turn, then describe the taxonomy and checklist produced from these sources.

The main conclusions from the literature review were on the different approaches to academic literacy and on the implications for testing. Of these main approaches (study-skills, genre-based or critically-socially-situated), it seems likely that the latter two are more authentic from a context-validity perspective, as they acknowledge the substantial differences that exist between subject and genre practices. Some tests do acknowledge this issue: admissions tests for specific subjects do exist for highly competitive subjects such as medicine, and this remains the most contextually-valid approach currently in existence.

However, this approach presents substantial practical difficulties, which must be carefully considered in the decision to choose this level of subject granularity. Firstly, it would be not only the exam papers that varied but also – to fully take into account the variations between subjects – clear and highly detailed guidelines for markers, and, potentially, different markschemes for each version. Such detailed documentation would require substantial research to ensure that variations were fully captured, and regular updating as the disciplinary requirements of each subject evolve. Secondly, quality assurance procedures must remain rigorous despite the fact that subject-specific papers will likely have a much smaller candidature: equivalence of construct and difficulty must be established across administration versions and, if the same institution uses subject-specific papers in more than one subject, equivalence of difficulty across subject versions. Also, for reasons of malpractice, new / limited exposure test material needs to be presented for each test administration. All of the above would drive up the cost on a per-candidate basis, which may lead to the exclusion of candidates from lower socio-economic backgrounds.

Importantly, however, even a subject-specific approach is in itself a compromise. Many proponents of the genre approach or the critically-socially-situated approach suggest that written norms occur at the institution, module or even tutor level. The argument as to whether this is appropriate for in-course university assessment is not relevant here, but I would strongly question whether such a highly granular approach is appropriate for an admissions context, where a candidate is likely to be taking multiple admissions exams with only a limited time to become familiar with the marking criteria. Therefore, the test designer has to resolve, not only the tension between validity, quality assurance and cost, but also the level of granularity that is fair to the candidates.

Given that compromise is inevitable, and given that my interest is in large-scale testing, I consider either subject-level or faculty-level grouping to be reasonable, with each having texts and topics relevant to each field. I would also recommend that this grouping include marking criteria and examiner training at the same level of granularity. A subject-level test is substantially more valid, but substantially more costly; further research is necessary to establish whether this extra validity is worth the many-fold increase in cost, which will be passed on to the candidate.

For the same practical reasons, the assessment context should focus on EGAP rather than ESAP.

An ALT should require the functions, lexis and syntax necessary for the academic environment; these are most distinct in reading, writing, and reading-into-writing activities. The essay is the most commonly used task type across university faculties, as a demonstration of and means of developing powers of informed and independent reasoning, while allowing independence of expression through choice of structure and arguments (Nesi and Gardner, 2012).

4.1.1 Comparison of the ICAS statement with the final taxonomy

I noted earlier that the ICAS statement was intended to be a list of skills required for undergraduate university study. As this list was the foundation of my taxonomy, I here present a comparison of ICAS with my final taxonomy and checklist, to gain an understanding of how complete this taxonomy is likely to be.

Table 6 shows those skills that were introduced into the taxonomy following review of the ICAS statement (that is, those subskills which could be considered as holes in the ICAS statement), and those subskills that are included in ICAS but not in the final checklist (those subskills in ICAS that can be considered less important). I remind the reader that the taxonomy is a complete list of all subskills collected, while the checklist is those subskills that scored over 16 and is intended for ALT analysis. The numbers in brackets represent the scores that each subskill received out of a possible total of 27, and thus the importance it was given by the three sources.

	In taxonomy but not ICAS statement	**In ICAS but not in final checklist**
Argument	Fully understand essay questions (22) Anticipate possible counter-claims (5)	None
Coherence and cohesion	None	Structure writing so that it moves beyond formulaic patterns that discourage critical examination of the topic and issues (12.5)
Academic language use	Discipline-specific writing (6) Use vocabulary precisely to produce the given effect (11) Text types: research proposals, dissertations, literature reviews (8)	Report facts or narrate events (11)
Eng. with sources	Reading strategies: skim, scan, read for detail (20) Understand inference (11) Understand and integrate quantitative data (10) Identifying suitable excerpts of text for direct / indirect quotation (10.5) Use of quotation, paraphrase and summaries to avoid plagiarism (17)	None

Table 6: subskills analysis for ICAS statement of competencies

4.1.2 Key findings from university websites

To provide a second viewpoint on the definition of academic literacy, the study-skills pages of three universities were chosen with the intention of achieving coverage of a variety of different institutions: the country in which they are based, a range of student entry qualifications and ages, and a variety of subject specialism. The three institutions chosen were as follows:

- The Open University, based in the UK and mainly catering to mature students, provides distance-learning courses only and therefore has an extensive list of resources dealing with study skills.
- Dartmouth College, based in Hanover in the US, is an Ivy League university with a strong focus on science and social science. The vast majority of the students are in the top 10% of their high-school class (Dartmouth, 2016).
- The University of Kent, based in the UK, is a mid-ranking university primarily made up of undergraduate students.

It is worth bearing mind that while all three websites were selected due to their detail and granularity (see section 3.1), of those three the University of Kent was the least detailed and the Open University the most, which is reflected in the variety of scores for each subskill.

4.1.2.1 Findings: University of Kent

Based on a survey of the categories on the study-skills pages, the University of Kent website (2017) has a stronger receptive-skills slant than the other two, focussing on two areas in particular: the first being note-taking from reading, specifically the use of summaries and paraphrasing as a means of comprehension. The other strong focus is on reading strategies: skimming, scanning and reading for detail are fully explained, and also the importance of contextualising the reading through surveying the text, through metatextual analysis and through comparison with the reader's world knowledge. Kent is the only website of the three that does not discuss variation of writing style between task type (register is only given a passing mention) or in fact any higher-level language skills except proofreading. The other key omission is the necessity of arguing with / critiquing a text; this idea is only given a passing mention in favour of fully comprehending a text.

Aside from the task types mentioned above, Kent has no other mention of specific task types, or of the use of quantitative data.

4.1.2.2 Findings: Open University

The Open University study-skills website (2017) has fairly comprehensive coverage of the subskills, with 28 of the 42 subskills as a key section or key idea. Like Kent, a key focus is reading skills, particularly strategies for interacting with the text such as note-taking and self-questioning to retain and understand key ideas. There is one page on practical techniques for dealing with difficult material – although this is for understanding complex ideas rather than complex language (it is intended for 1st language speakers rather than directly targeted at 2nd language speakers).

However, reading is usually a key idea (scoring 3) rather than a key section (scoring 4), while aspects of writing often receive pages to themselves (scoring 4). Five of the highest-scoring skills relate to coherence and cohesion, while four are on variety of writing purposes, audiences and task types. Academic-style English is also significant, dealing with register, vocabulary and some grammar – broken down into tenses, voice, key verbs, nominalisation, and common errors. The study-skills section links to another website that deals with language skills in greater detail, but this website is not included in this analysis: the focus of this other website is general English proficiency (rather than EAP) and so describes language features in great detail, such as grammar points. LP is included in academic literacy, and I did not begin this research with any preconceptions as to how important the three sources consulted would consider LP. However, this was the only one of the three websites, or in fact of any of the sources consulted, which covered language features in detail. This is in line with my definition of academic literacy as a skill to be developed in L1 and L2 speakers alike, and thus a construct that includes but also goes beyond LP (see section 2.1.2). As no other sources dealt with LP, inclusion of this LP website would not have resulted in a change in the results for RQ1.

The Open University has very few subskills that are not covered (four not mentioned at all, and two implied). Evaluation of texts is dealt with at the undergraduate level only in terms of evaluating if the texts are appropriate for the reader's purpose. Critical analysis of the quality of the source appears in the postgraduate section and therefore received a score of zero for the purpose of this study. Analysis of text structure is dealt with in terms of headings, contents pages and other areas outside the body of the text.

4.1.2.3 Findings: Dartmouth College

The Dartmouth College website (2017) has less of a receptive-skills slant than the other two; those aspects of reading dealt with are often under the umbrella of the purpose of reading, including types of reading (skimming, etc) and also reading speed. Active reading is a key idea across all elements of Dartmouth's study-skills / reading / writing pages, the website stating that to read passively is to 'hold off making any intellectual response to the text until after you've finished reading it.' and that to read actively is to 'enter the conversation' in that academic discipline. Like the Open University, reading skills are key ideas (scoring 3) rather than key sections (scoring 4). The key sections are often devoted to writing skills, especially materials to support first-year writers. Structure is a key idea in these pages, specifically the purpose of structure in supporting and clarifying the writer's argument. The webpages also discuss thesis statements in much greater depth than the other websites in this literature review, as well as discussion of synthesis of sources. It also deals with criticism, deconstruction and reader-response. Redrafting is seen as a key part of the writing process, with a series of checklists for content covering introduction, thesis, structure, paragraphs, argument and logic, and conclusion; for theses, finding the overall idea or unpacking assumptions and generating counter-claims.

Like the Open University, Dartmouth also highlights the fact that disciplines vary in terms of their expectations and states this the most explicitly of the three websites: 'each of the disciplines has its own way of constructing knowledge, of organising that knowledge, of using evidence, and of communicating within the field' (University of Dartmouth, 2017).

For Dartmouth, the score that is the most different from the others is the subskill 'use of quotation, paraphrase and summaries to avoid plagiarism', which is respectively a key idea and a key section for Kent and the Open University, but is not mentioned on the Dartmouth website at all.

4.1.2.4 Websites: overall trends and additions to the taxonomy

Table 7 summarises the ratings for the top-ten-rated subskills for websites. Seven of these ten relate to writing, with two for reading and one for critical evaluation of texts.

	Website total
Fully understand essay questions	12
Develop main point or thesis	11
Proofread to eliminate errors in grammar, mechanics and spelling, using standard English conventions	11
Structure writing so that it is clearly organized, logically developed and coherent	10
Reading strategies: skim, scan, read for detail	10
Organize information at both a section and paragraph level	10
Use revision techniques to improve focus, support and organization	10
Read texts of complexity without instruction and guidance	9
Critically assess the authority and value of research materials that have been located	9
Develop thesis convincingly with well-chosen examples, reasons and logic	9

Table 7: Top ten scoring subskills for university websites

Key section= 4; Key idea = 3; Mentioned = 2; Implicit = 1; No content = 0

In terms of task types, the University of Kent adds to the ICAS list with dissertations, research proposals and literature reviews, but makes no mention of laboratory reports. Neither the Open University or Dartmouth College add any new task types to the list.

Table 8 shows the subskills that were added to the taxonomy at this stage.

Argument	Fully understand essay questions Draw conclusions from given reading
C&C	None
Academic language use	Text types: research proposals, dissertations, literature reviews Discipline-specific writing Use vocabulary precisely to produce the given effect
Engagement with sources	Use of quotation, paraphrase and summaries to avoid plagiarism Selecting appropriate texts Understand and integrate quantitative data Identifying suitable excerpts of text for direct / indirect quotation Understand inference Reading strategies: skim, scan, read for detail

Table 8: Subskills added to the taxonomy (included in websites but not in ICAS)

I have now reviewed the findings from the websites in light of the literature. Next, I will describe the key findings from the ALTs.

4.1.3 Key findings from existing academic literacy tests

4.1.3.1 Test of English for Educational Purposes, University of Reading

The University of Reading's TEEP (University of Reading, 2017) is an integrated-skill reading-and-listening-into-writing test of academic literacy. It is accepted by many UK universities as proof of English-language ability at both undergraduate and postgraduate level, and is also used as an end-of-course summative assessment.

It consists of six parts, of which parts 3-5 are relevant to this thesis. Parts 3-5 are thematically linked: parts 3 and 4 provide input for an essay to be written in part 5. Part 3 uses objective question types to test general and detailed comprehension, including academic English writing structures and inference, of one academic-style text of approximately 1000 words. In part 5 candidates are required to draw ideas from the reading texts and the lecture extract which forms the part 4 listening component, to produce a 350-word essay task which presents the source material as well as expressing and defending their own ideas. The writing is marked on content, including referencing of sources, full exploration of ideas and relevance to the essay question; argument and organisation, covering comprehensibility, quality of argument and structure; and grammar and vocabulary, covering range, accuracy, appropriacy and register; there is also a holistic, overall impression, mark.

While consisting of a structure that presents practical difficulties for large-scale test production, the TEEP covers nearly all aspects of the reading-into-writing construct presented by ICAS. Three of the subskills are only partially covered: summarising information is to some extent covered by the requirement to refer to sources, although it would be possible to answer the essay at a lower level using quotation only. Three further areas are not covered at all, as can be expected in a cross-discipline exam: alternative text types, such as research proposals or laboratory reports, are not required, and therefore neither is the function of reporting facts or narrating events in writing up primary research (e.g. results of an experiment). Discipline-specificity, with all the associated issues of discipline-expectations, is also not tested.

4.1.3.2 NYSTCE: New York State

The NYSTE is required for all teachers in New York State, including English L1 speakers, and is taken by graduate students only (note that this purpose is different from that of many ALTs: the majority of the latter are used for university entrance). The exam is made up of three papers, one of which is the Academic Literacy Skills Test (ALST).

Task type 1: Selected-response reading comprehension

The ALST Test Design and Framework performance indicators (New York State Education Department, 2014) indicate that this component covers both literal comprehension of the texts and figurative understanding, this latter including inference, authorial attitude and drawing conclusions from given information. There is also an element of language awareness, in that questions can target 'how specific word choices shape meaning and tone in a text' (ibid).

This test is the only one of the three ALTs in this thesis that includes extensive reading texts at two points: once in the reading comprehension paper, and again as part of the reading-into-writing paper, which seems like a duplication of content. Certainly, the purpose of this exam (teacher certification rather than university entrance) means that the construct targeted is substantially different: while the concern with university entrance is to mirror university activities, it could be argued specific functions required for teacher certification can be specifically targeted by the multiple-choice format, or that some of the items targeting argumentation or authorial attitude may be more difficult to elicit in a reading-into-writing format, such as in the following stems:

> 'Which of the following statements, if added to the passage, would weaken the author's statement that Amelia Earhart's whereabouts are unknown?'
>
> 'In which of the following excerpts does the author most clearly express disbelief?'
>
> 'Which of the following statements from the paragraph is most convincingly an opinion as opposed to a fact?'
>
> Postman, 2015, pp 159-161

Task type 2: Writing from sources

There is no crossover in terms of topic between the reading and writing components.

The first part of this paper requires the candidate to summarise three sources from a particular perspective and then compare and contrast the viewpoints of each. It requires examples from the sources to be selected and presented as part of the argument. Postman gives an example question as '…review each passage and compare and contrast the views each takes on raising the United States debt limit' (Postman, 2015, p. 195).

The second part of the paper is a 400-600-word persuasive essay. The instructions include the requirements to 'demonstrate that you understand the topic, use logical reasoning to expand and extend the points made in the passages, provide evidence from all three sources to support your claim, including the graphic' and, 'present and refute a counterclaim' (Postman, 2015, p.196).

These instructions closely mirror the priorities of the taxonomy drawn up thus far in the focus on critical engagement and argumentation. LP is a minor point in the instructions given to candidates, these instructions highlighting clarity and coherence and organisation of complex ideas, but also specifying a formal register and 'the standards of written English grammar, usage and punctuation (Postman, 2015, p.192).

4.1.3.3 Tests: additions to the taxonomy and overall trends

Subskills	NYSTE (5)	TEEP (5)	Website total (12)
Fully understand essay questions	5	5	12
Reading strategies: skim, scan, read for detail	5	5	10
Understand inference	5	5	1
Understand and integrate quantitative data	5	0	5
Identifying suitable excerpts of text for direct / indirect quotation	5	2.5	3
Comparing and contrasting two (or more) texts	5	0	0
Draw conclusions from given reading	5	5	3
Use of quotation, paraphrase and summaries to avoid plagiarism	5	5	7
Use vocabulary precisely to produce the given effect	5	5	1
Anticipate possible counter-claims	5	2.5	0

Table 9: subskills in tests but not the ICAS statement

Table 9 shows the subskills that are tested by the NYSTE and/or TEEP, which are not included in the ICAS statement (the numbers given in brackets are the total possible score). Of the skills listed, two do not appear in either the ICAS statement or the websites: comparing and contrasting two (or more) texts, and anticipating possible counter-claims. It is interesting to note that these subskills are primarily receptive (7 subskills), with only one subskill that can be considered writing only (precise use of vocabulary).

	NYSTE (5)	TEEP (5)	Website total (12)
Generate ideas for writing by using texts in addition to past experience or observations	5	5	6
Understand inference	5	5	1
Decipher the meaning of vocabulary from the context	5	5	3
Identify authorial attitude	5	2.5	4
Identify the evidence which supports, confutes, or contradicts a thesis	5	5	4
Identifying suitable excerpts of text for direct / indirect quotation	5	2.5	3
Selecting appropriate texts	0	0	6
Understand 'rules' of various genres	2.5	5	4
Draw conclusions from given reading	5	5	3
Use vocabulary precisely to produce the given effect	5	5	1
Text types: research proposals, dissertations, literature reviews	0	0	8

Table 10: subskills with a large difference between test and website coverage

As there was a very limited range of scores given, it is not meaningful in this case to discuss the top-ten-scoring subskills (for example, 25 of the subskills received the maximum score of 10) Instead, Table 10 lists the subskills with large differences between test and website coverage, that is, where one source has placed much greater weight on a subskill than the other sources. In these cases, it is mostly the case that the test has more complete coverage than the websites, with nine subskills where the test has greater coverage versus two where the websites' coverage is greater.

Four subskills appear in both Table 9 and Table 10: they appear in the tests but not in ICAS and receive a low score in website coverage; those skills are understanding inference, identifying suitable excerpts of text for direct / indirect quotation, drawing conclusions from given reading and using vocabulary precisely to produce the given effect.

Following analysis of the two ALTs, the two subskills that had not been captured so far by ICAS or the websites were added to the taxonomy. Those subskills were comparing and contrasting two (or more) texts, and anticipating possible counter-claims.

4.1.4 Final taxonomy

The overall taxonomy comprised all the subskills from the ICAS taxonomy, the websites and the tests, with subskill scores ranging from 26 (most prominent) to 5 (least prominent).

No additions were made to the taxonomy as a result of the expert judgement stage (see section 3.1.3). The following subskills were combined as they targeted very similar skills:

Skills in taxonomy with overlap	Skill(s) appearing in final checklist
• Read texts of complexity without instruction and guidance • Reading strategies: skim, scan, read for detail • Decipher the meaning of vocabulary from the context	Reading strategies: skim, scan, read for detail Decipher the meaning of vocabulary from the context
• Synthesize information in discussion and written arguments	Synthesize ideas from several sources

• Synthesize information from assigned reading • Synthesize information from reading and incorporate it into a writing assignment	
• Use of quotation, paraphrase and summaries to avoid plagiarism • Summarise information • Summarize ideas and/or information contained in a text	Use of quotation, paraphrase and summaries to avoid plagiarism

Table 11: Synthesised subskills following collection

After consultation with the experts, the subskill 'organise information' was altered to read 'organise information at a section / paragraph level' to distinguish it from 'structure writing so that it is clearly organised, logically developed and coherent'.

4.1.5 Creating the checklist and setting cut scores

At this stage, my opinion and that of the experts consulted was that not all of the subskills in the overall taxonomy were necessary in an ALT. Therefore, a shorter checklist of subskills was created that included the highest-scoring (and therefore most prominent) subskills.

Two possible cutoff scores were considered and the resulting checklists also submitted to the experts above for comment. The first cutoff score considered was 20 (that is, only subskills scoring 20 or above were included) as this score indicates the particular subskill was tested in both of the tests and treated as a key idea in at least two of the university websites. Subsequently, however, the cutoff was lowered to 16 based on two considerations: The first, that the experts consulted above suggested that the range 16-20 contained several subskills that were key areas in academic performance (such as ‘summarize ideas and/or information contained in a text’, scoring 19.5, and ‘use of quotation, paraphrase and summaries to avoid plagiarism’, scoring 17), but that this was not the case for subskills scoring below 16; the second, that while the subskills with scores above 16 received scores that were very close together, often separated by as little as half a point, there was a clear gap between the subskills that scored 16 and the next lowest score of 13 (see figure 2 below).

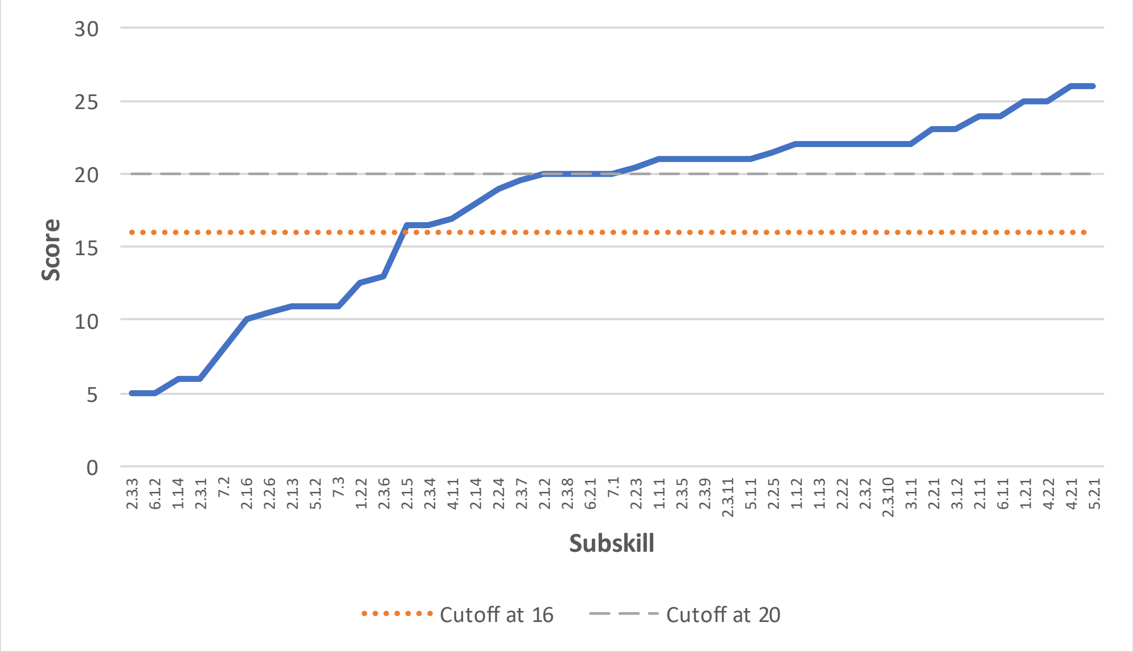

Figure 2: Scores for each subskill, arranged in ascending order

The final taxonomy (appendix 8.1) consisted of 42 subskills, and the checklist of 25. The checklist appears on the following pages. Crossed out text indicates that this subskill was changed following RQ2 (see section 6).

Cognitive process	**Competency number**	**Academic literacy sub-skills**	**Total for competency**
Conceptualisation: task representation and macro-planning	1.1.1	Generate ideas for writing by using texts in addition to past experience or observations	21
	1.1.2	~~Fully understand essay questions~~ Fully understand the task requirements of essay questions, including understanding of genre conventions, readership and wording of the rubric / task	22
	1.1.3	Duly consider audience and purpose	22
Conceptualisation: revising macro plan	1.2.1	Structure writing so that it is clearly organized, logically developed and coherent	25
Meaning and discourse construction: careful global reading	2.1.2	Reading strategies: skim, scan, read for detail	20
	2.1.4	Decipher the meaning of vocabulary from the context	18
	2.1.5	Identify authorial attitude	16.5
Meaning and discourse construction: selecting relevant ideas	2.2.1	Identify the main thesis of a whole text	23
	2.2.2	Determine major and subordinate ideas in a particular passage	22
	2.2.4	Identify the evidence which supports, confutes, or contradicts a thesis	19
	2.2.5	Critically assess the authority and value of research materials that have been located	21.5
Meaning and discourse construction: connecting and generating	2.3.2	Make connections to related topics, information or prior knowledge, even when they are not obvious.	22
	2.3.4	Understand 'rules' of various genres	16.5

	2.3.5	~~Understand separate ideas and then be able to see how these ideas form a whole~~ Understand separate ideas within one source and see how these ideas form a whole	21
	2.3.8	~~Synthesize information from several sources and incorporate it into a writing assignment~~ Understand separate ideas from several sources and see how these ideas form a whole	24
Translation	3.1.1	Vary sentence structures and word choice as appropriate for audience and purpose	22
	3.1.2	Use vocabulary appropriate to college-level work and the discipline	23
Organising ideas in relation to input texts	4.1.1	Use of quotation, paraphrase and summaries to avoid plagiarism	17
Organising ideas in relation to writer's own texts	4.2.1	Develop main point or thesis	26
	4.2.2	Organize information at both a section and paragraph level	25
Low-level monitoring and revising: editing while writing	5.1.1	Link ideas appropriately	21
Low-level monitoring and revising: editing after writing	5.2.1	Proofread to eliminate errors in grammar, mechanics and spelling, using standard English conventions	26
High-level monitoring and revising: editing while writing	6.1.1	Develop thesis convincingly with well-chosen examples, reasons and logic	24
High-level monitoring and revising: editing after writing	6.2.1	Use revision techniques to improve focus, support and organization	20
(Task types)	7.1	Provide essays	20

Table 12: Checklist of subskills necessary for ALTs

4.2 RQ1b: initial results

To explore and begin to validate the academic literacy construct drawn up in RQ1a, I analysed an instance of an ALT: the CAEALT (see section 3.2.1 for a description of this test).

RQ1b: To what extent is the CAEALT representative of the construct defined in RQ1a?

I analysed the CAEALT against the checklist (see section 3.1 for more information on the procedure for this). This analysis took place at two points: first, immediately following the construction of the checklist as above; second, following the RQ2 results, as the case studies included in these results brought further insight into the test's coverage of the checklist (see section 6.2). Of these analyses, the second should be considered the most representative of the test, but the initial analysis is presented here as context for the reader in the upcoming description of RQ2. This analysis also forms appendix 8.2).

4.2.1 First iteration of RQ1b results

The first analysis of the CAEALT was based on my expert judgement rather than on candidate performance (analysis based on candidate performance can be found in section 6.2).

According to this first analysis, almost all the 25 subskills on the checklist are covered: 23 were categorised as required for successful completion of the task (scoring 5), two as possibly required (scoring 2.5), and none as not required (scoring 0).

Table 13 shows the subskills that are only partially or not included.

Subskills	CAEALT (5)	NYSTE (5)	TEEP (5)	Website total (12)
Understand 'rules' of various genres	2.5	2.5	2.5	4
Use revision techniques to improve focus, support and organization	2.5	2.5	2.5	10

Table 13: coverage of subskills partially or not covered by the CE ALT

While, in the exclusion of these subskills, the CAEALT is in agreement with the other ALTs, in the first case the subskill was not explicit on university websites, but for the second, it was. This is likely to reflect the fact that large-scale revision of scripts cannot take place within an exam setting.

4.3 RQ1: Discussion

The methodology chapter for RQ1 laid out the process by which I arrived at a taxonomy of academic literacy subskills: First, the literature review suggested an ALT should cover EGAP, focussing on reading-into-writing at a study-skills or faculty-specific level; sections taken from the ICAS statement of competencies formed the basis of the checklist; these sections were compared against the study-skills pages of three university websites and the constructs of two ALTs. These three sources together made up the academic literacy taxonomy; the higher-scoring skills on this taxonomy became the academic literacy checklist (Table 12).

This discussion will compare the relative importance of different subskills and categories of subskills across the sources, comparing each source with both the taxonomy and the checklist, before a general discussion of issues surrounding the creation of this taxonomy.

4.3.1 ICAS statement: relationship with the websites and tests

The literature on academic literacy contributed to the results of RQ1 in two ways: first, in terms of providing general guidelines (a focus on reading and writing over speaking and listening; grouping subjects by faculty; using the essay task type; a focus on academic English for General rather than Specialised Purposes), and second in terms of providing the initial list of subskills (the ICAS statement of competencies).

Section 4.1.1 compared the ICAS statement with the final checklist (see Table 6). That there are few additions to the ICAS statement, which is to say, that the ICAS statement and the final taxonomy are similar, is entirely to be expected, as ICAS was the result of surveys of faculty members (presumably faculty members are also responsible for the foci of university websites). Those subskills that were not included in ICAS, but were in the taxonomy, fall into four categories:

- Faculty-specific subskills (such as 'discipline-specific writing' or 'quantitative data') faculty-specific requirements were out of scope for ICAS.
- Not necessary until the end of an undergraduate degree (such as 'research proposals, dissertations, literature reviews'). ICAS focusses on skills required for undergraduate entry.
- Skills that have a slight change in focus between ICAS and other sources (such as 'use vocabulary precisely to produce the given effect', which is only slightly different from 'use vocabulary appropriate to college-level work and the discipline' (ICAS, p.39)). Both appeared in the taxonomy, as strictly one concerns lexical accuracy rather than lexical range/register, but there is definite crossover between these two subskills.

- More specific instances of a skill discussed in ICAS in a more general way (such as 'use of quotation, paraphrase and summaries to avoid plagiarism', which is best reflected in ICAS as 'correctly document research materials to avoid plagiarism' (ibid, p.41)).

Of these additions, two more surprising non-inclusions in the ICAS taxonomy should be particularly noted. First, I noted above that the subskills 'use of quotation, paraphrase and summaries to avoid plagiarism' and 'identifying suitable excerpts of text for direct / indirect quotation' are not clearly reflected in the ICAS statement, which has subskills for 'correctly document research materials to avoid plagiarism' and 'synthesize information from reading and incorporate it into a writing assignment'. The suggestion to be deduced from this non-inclusion is that students pre-university are required to have a basic understanding of the concept of plagiarism, and be able to include ideas from others in their writing, but are not required to do this in a structured way which fully represents and documents the authors of those ideas. Whether this is in fact the case is not clearly reflected in the literature, and may be in need of further research.

The second surprising non-inclusion is 'anticipate possible counter-claims'. ICAS has a critical-thinking category of subskills, which includes 'compare and contrast own ideas with others', 'interrogate own beliefs' and 'identify evidence which supports, confutes, or contradicts a thesis', but does not directly target the subskill of acknowledging and counteracting the arguments of others in the student's own writing. Related subskills do appear in the final checklist, including identification of evidence that contradicts a thesis and supporting a thesis with well-chosen reasons and logic. Although I would argue that anticipating counter-claims is a distinct subskill, and a necessary skill in the academic world, perhaps it has in fact been subsumed into other subskills by other writers.

It is reassuring that there are only two subskills that were included in the relevant sections of ICAS but not in the final checklist, which suggests that the priorities of the two lists are similar. It is to be expected that 'report facts or narrate events' is not included, as my literature review suggests that essays are the most-used task type. However, I question whether 'structure writing so that it moves beyond formulaic patterns that discourage critical examination of the topic and issues' may not be necessary for university success: a structure that is tailored to the particular argument being presented, rather than a structure following a pre-set template, seems key to the argument being truly effective. This is an idea I will return to in section 6.1.3.

4.3.2 Websites: relationship with the literature and tests

In general, the view of academic literary presented by the websites is more granular than for ICAS, with a wider variety of genres discussed. The subskills given in the section above are a good example: while websites specifically require use of quotation, paraphrase and summary to avoid plagiarism, ICAS only mentioned correct documentation of research materials.

Section 4.1.4 noted that there were 12 subskills with a variety of score of three or more across the three websites, and which can therefore be considered more controversial in terms of inclusion in a list of academic literacy subskills. In most of the cases, this is due to the Open University scoring substantially more than the other two.

Table 14 below shows those subskills which were added to the taxonomy at the website stage.

	Subskills included in websites but not in ICAS
Argument	Fully understand essay questions ~~Draw conclusions from given reading~~
C&C	None
Academic language use	~~Text types: research proposals, dissertations, literature reviews~~ ~~Discipline-specific writing~~ ~~Use vocabulary precisely to produce the given effect~~
Engagement with sources	Use of quotation, paraphrase and summaries to avoid plagiarism ~~Selecting appropriate texts~~ ~~Understand and integrate quantitative data~~ ~~Identifying suitable excerpts of text for direct / indirect quotation~~ ~~Understand inference~~ Reading strategies: skim, scan, read for detail

Table 14: subskills added to the taxonomy from websites
Note: Struckthrough text indicates that this subskill did not appear on the final checklist

Reading strategies is an interesting omission from ICAS, as the ability to interact with texts in a purposeful way would seem to be necessary to deal with the quantity of reading necessary at university. As with some of the subskills above, the implication is that this is a skill to be acquired during undergraduate study.

4.3.3 Tests: relationship with the literature and websites

The view of academic literacy presented by the tests is in line with the guidelines created at the end of the literature review, with the exception of grouping by faculty, which I have recommended as a potential compromise between the study-skills and genre-based models. I have raised concerns how well the tests would deal with lifting of input material and input language transformation, and whether listening-into-writing is an appropriate integrated task.

In terms of subskills, while the tests are much more granular and explicit in their approach to subskills, there are few subskills that are not covered, and none are a cause for concern. I also note that there is less variation between the approach of each tests than there was in the case of websites.

The two tests analysed for this study serve slightly different constructs: only the Reading TEEP is for university entry (and thus a direct parallel can be drawn with ICAS), and is aimed at L2 students. The other test, the NYSTE, is aimed at graduates, and is part of the teacher certification process for both L1 and L2 students. Both exams are intended to test LP skills as well as academic literacy skills.

The extent to which the requirements for an ALT as per the literature are covered is given in table 13.

	TEEP	NYSTE
EGAP	Y	Y
Grouped by faculty	N	N
Cover the functions, lexis, syntax necessary for academic environment	Partial	Partial
Integrated skills	Y	Y
Reading into writing	Y (and listening)	Y
Essay task	Y (and other writing tasks)	Y

Table 15: AL test coverage of issues raised by the literature review

Neither test has chosen to provide alternative versions by faculty, presumably for reasons of practicality. The literature review discussed the arguments for subject- or faculty-specific versions: namely, that there are key differences in writing practices between subjects: format, rhetorical style, metadiscourse and lexis can all vary. However, I also note that, unless marking criteria, task type and examiner training vary, the extent to which a paper can be said to be tailored to a faculty is limited. Therefore, the lack of grouping by faculty in these two tests may not be particularly meaningful.

Three other general concerns were raised by the literature review (section 2.1.3). First, extensive lifting from source materials can be problematic, which can be mitigated by giving the candidates restrictions on lifting 'as in real-life rules concerned with plagiarism' (Khalifa and Weir, 2009, p.91). Second, candidate scripts can suffer from a lack of development of the ideas given in the source texts, to be mitigated by requiring 'input language transformation' (ibid). Finally, real-life writing relies on multiple extended secondary sources, which is impractical under exam conditions.

Neither test explicitly addresses the first concern in the task rubric, and therefore it would be interesting to see how much lifting takes place in these exams (see also sections 2.1.3 and 6.1.3). I cannot comment on the possibility of underdeveloped ideas in candidate scripts as I do not have access to these. However, if this is indeed an issue in scripts, this is less problematic in the NYSTE, as this skill is not likely to be as necessary in the teaching sphere as in the academic.

Of the subskills that appeared in the tests but not in the ICAS statement, only two scored highly in both the tests and the websites: fully understand essay questions (10 for tests, 12 for websites) and reading strategies (10, 10). That two sources agree suggests that these should be included in a definition of academic literacy.

There were several subskills which were included in tests, but which were not included in ICAS and scored low in websites:

- Understand inference
- Identifying suitable excerpts of text for direct / indirect quotation
- Draw conclusions from given reading
- Use vocabulary precisely to produce the given effect

While the low scores for ICAS and websites may suggest the inclusion of these subskills in tests is not justified, in fact these are highly specific subskills, as is necessary for a mark scheme or construct, but such specificity is not necessarily required in the other two contexts.

The first two of these are particularly interesting as they scored highly in the AL test analysis. An argument could be made that these are skills not often explicitly taught to students, and thus will not appear on university websites.

4.3.4 Subskills with unexpected overall scores

Throughout this discussion, there have been a few subskills with more unexpected scores. First, given the prevalence of quantitative data in ALTs and academic-focused LP tests, it may be surprising that understanding and integrating quantitative data was one of the lower scoring subskills in the table, with a total score across all sources of 10: inclusion in the NYSTE, a website score of 5, and no mention in ICAS. It seems likely that this is a faculty-specific skill: students in the humanities would not require this, and data presentation conventions across other faculties vary considerably – a physical scientist would query the lack of error bars in such tests, for example. For this reason, I suggest that quantitative data is a key part of subject-specific knowledge, and thus not included in EGAP nor the final checklist.

I mentioned above that the subskill 'anticipate possible counter-claims' was not included in ICAS. In fact, this was not a subskill emphasised by the websites either, scoring 0, and the related subskill 'identify evidence which supports, confutes, or contradicts a thesis' only scored 4 across the websites. This seems surprising, as it would appear to be an important academic skill. There are three possible reasons for this: that critical engagement is not prioritised until towards the end of an undergraduate degree; that it is usually taught in subject-specific classes (and therefore is not included in general study-skills information); that it is not a skill that students struggle with (and therefore there is no need for it to be specifically targeted). This last seems unlikely.

Finally, I note that appropriate use of quotation, paraphrase and summaries to avoid plagiarism also scored relatively low, considering the importance that is often placed on this in EAP courses: It had no mention in ICAS and scored 7 for websites; the bulk of its points came from the ALTs, where it received the maximum score. While this may on the surface seem unusual, this is in fact a consequence of it not being mentioned by Dartmouth at all, which may be in line with its profile as a science-leaning college, and the website's strong focus on receptive skills: this subskill was important to the other two universities, scoring 4 and 3.

In summary, in this section I have discussed several areas of interest. First, the sources consulted suggest that the following skills, which initially may appear to be fairly basic skills, may in fact be acquired at undergraduate level: correct use of sources to avoid plagiarism; synthesis of sources; anticipate possible counter-claims.

Next, there seems to be some lack of agreement as to which subskills can be considered faculty specific and which are not. For example, understanding of quantitative data is valued by study-skills format tests, but, as discussed above, there is an argument that this subskill is in fact faculty specific as both the quantity of quantitative data and the format this data appears in will vary substantially across subjects. I have also hypothesised that 'anticipate possible counter claims' may be a subskill that is taught in faculty-specific classes, when it might seem that this is a more general skill.

I have suggested that reading strategies: skim, scan, read for detail; understand inference; identify suitable excerpts of text for direct/indirect quotation, are all areas that may not be explicitly taught, but which are targeted in tests. If this is the case, and inclusion of these subskills in the tests is justified, it may be that they would benefit from being explicitly taught.

Finally, I have questioned whether 'structure writing so that it moves beyond formulaic patterns that discourage critical examination of the topic and issues', which is included in ICAS but not significantly covered in the other two sources, is in fact a useful academic subskill that should be both explicitly taught, and a way found to include this subskill in ALTs.

The second part of RQ1 is the extent to which the CAEALT is representative of the checklist here arrived at. An initial evaluation of this test was made in section 4.2.1. However, as this evaluation was revised during the process of RQ2, this will instead be discussed in section 6.2.

5 RQ2: Results and discussion

RQ2: What is the relationship between performance in the CAEALT and academic success?

5.1 RQ2: Quantitative results

This section will discuss trends among the CAEALT marks given and variety across university grades, before giving correlations between university grades and other measures of ability.

5.1.1 Quantitative analysis

5.1.1.1 Data

An overview of the results appears in Table 16. The full results appear in appendix 8.5.

	Average university grades	Average overall mark	Average argument	Average C&C	Average Acad. Lang. use	Average Eng. with sources
Undergraduates	63.19	4.7	5.00	4.50	5.10	4.10
Postgraduates	75.73	5.71	5.71	5.57	5.85	5.71

Table 16: CAEALT means

When considered by ranking, the postgraduate students are entirely in ranks 1-4 of scores awarded (there was a large number of rankings shared between 2 or more candidates), with an average overall mark of 5.71 and no individual category awarded below a 5 (of a maximum mark of 6). Among these students the strongest skill was academic language use (5.85 average) and the weakest coherence and cohesion (5.57 average), the other two categories scoring 5.71 average each.

Among the undergraduate students the average mark was a whole band lower at 4.7. As with the postgraduate students, academic language use was the strongest category (5.1 average). Notably, coherence and cohesion scored below the overall undergraduate average at 4.5 (slightly over a band lower than the postgraduates), and the weakest category was engagement with sources (4.1 average, slightly over a band and a half lower than the postgraduate students).

University grades are listed in appendix 8.6, and averages in Table 16. The number of grades given varied significantly: two of the postgraduate students reported one grade only (the overall grade for their non-modular qualification) while some had as many as 11 separate grades. All of the undergraduate students were enrolled in modular assessment and therefore had between 2 and 8 grades for the year. Taking both cohorts together, the standard deviation (not including the two students with one grade) was 4.35 and the average number of grades reported was 4.95. In terms of the undergraduate versus the postgraduate averages, I note that the postgraduate grade average is over ten marks higher than the undergraduate. This is not surprising, as to some extent a good undergraduate score is necessary to enrol on a postgraduate degree and therefore those on postgraduate courses would have been at the higher end of their undergraduate cohorts. Additionally, it is worth mentioning that the pass mark on postgraduate courses is often higher than for undergraduate, at 50% rather than 40%.

The sample's grades did not cover a wide range of abilities: they were almost entirely 2.1 and above i.e. the more successful students in a cohort, with no students receiving a 2.2 and one only receiving a 3rd. It should be noted here that equivalence of grades has not been demonstrated: grades are from a range of universities and from a range of examiners. For reasons discussed in section 2.2, this is an unavoidable issue when using opportunity sampling.

Self- and tutor assessments are listed in appendices 8.7 and.8.8.

The self- and tutor assessment results were considered first as raw scores. However, as no standardisation took place in the use of the scales, there was little variation in mean score between candidates (with an average of 4.13, standard deviation of 0.77 for self-assessment and an average of 4.40 and standard deviation of 0.74 for tutors), despite differing grades and ALT results. Self- and tutor assessment results were then normalised to focus the analysis on variation in scores.

It was only possible to get tutor assessments for a small number of the candidates (n = 5). This was for a few reasons: first, many of the postgraduate students were no longer in contact with their tutors, or were part of large cohorts where the tutor would have limited knowledge of a particular student's performance. Second, a few tutors had concerns over confidentiality which could not be allayed by discussing the consent given by participants. Finally, due to time constraints the second half of results collection took place at the same time as the nationwide university staff strike, meaning that some participants' tutors did not have the time to provide an assessment. The small number of tutor results available means that the tutor assessment results reported below must be extremely tentative.

5.1.1.2 Analysis

Correlations were calculated as follows: for overall correlations, grades, overall tutor assessment and overall self-assessment were correlated with the overall CAEALT score. For each of the four separate categories (argument, coherence and cohesion, academic language use and engagement with sources), university grades and the subskills relating to each category in the tutor and self-assessments were correlated with the CAEALT score in each category.

5.1.1.3 Grades to ALT

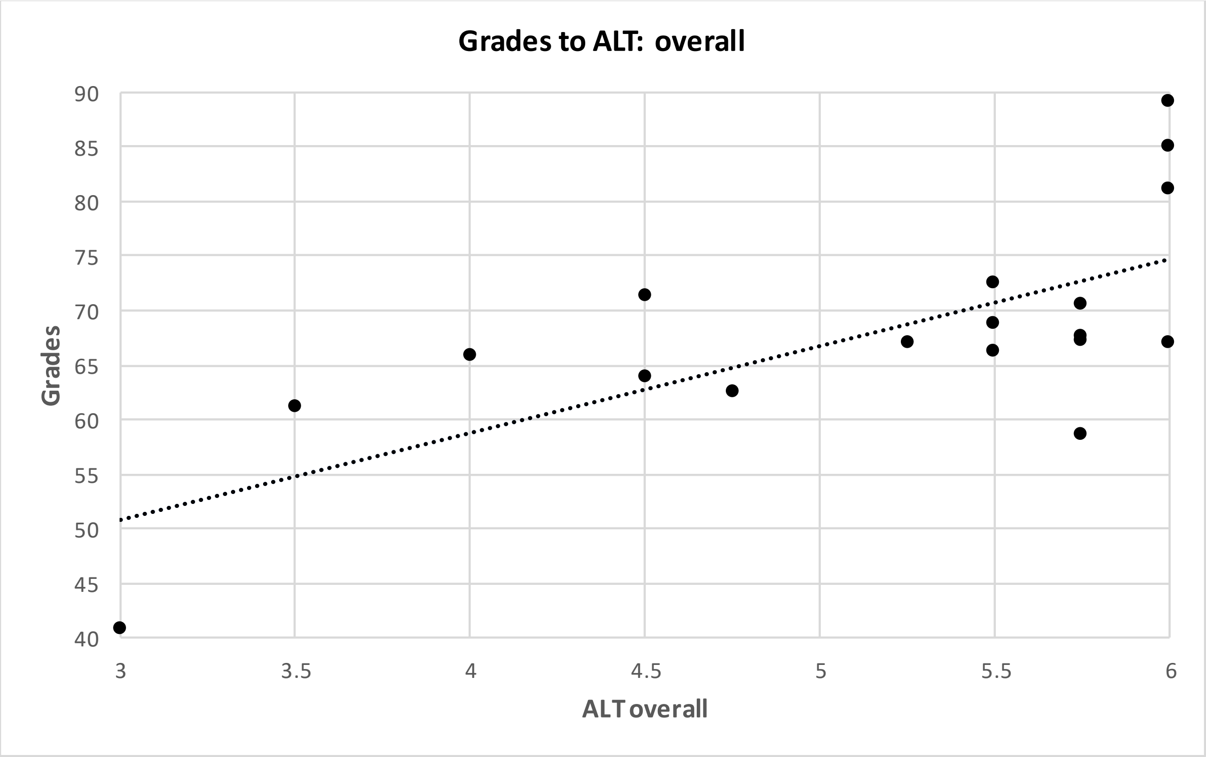

Figure 3: Grades to CAEALT, overall

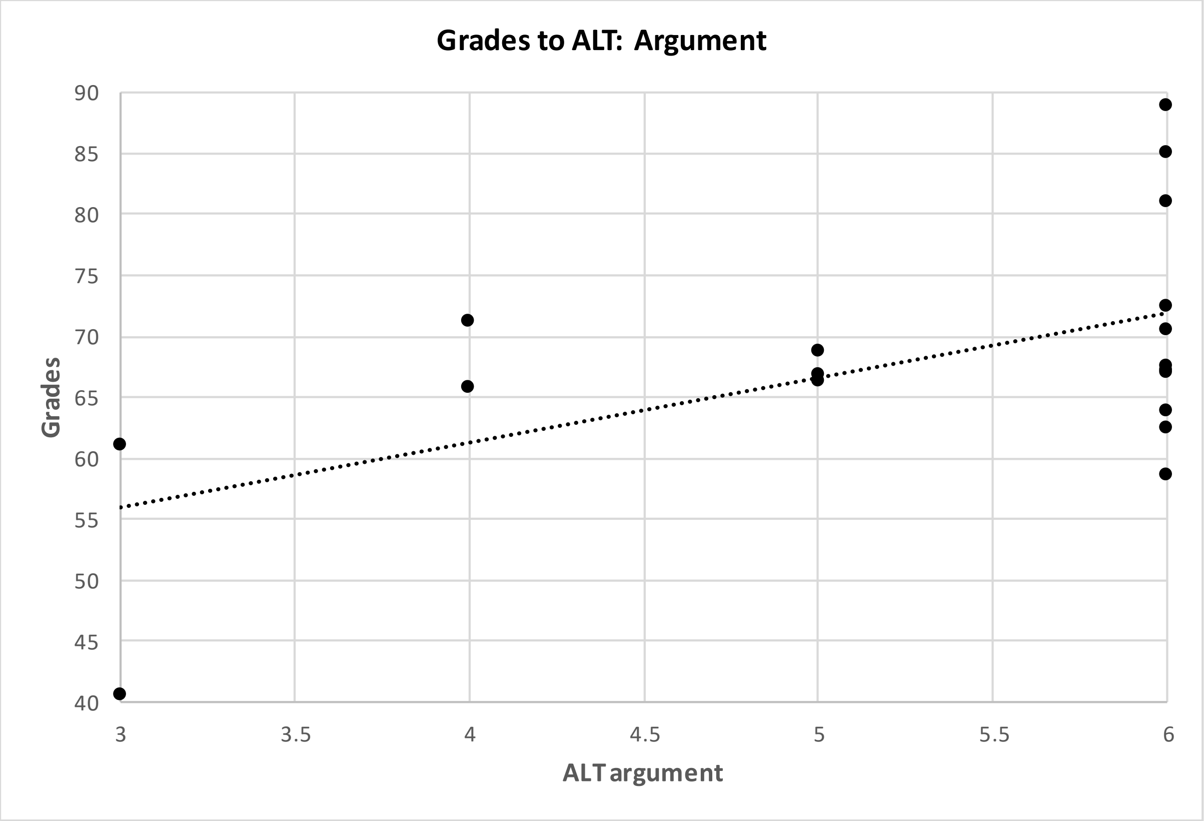

Figure 4: Grades to CAEALT, argument

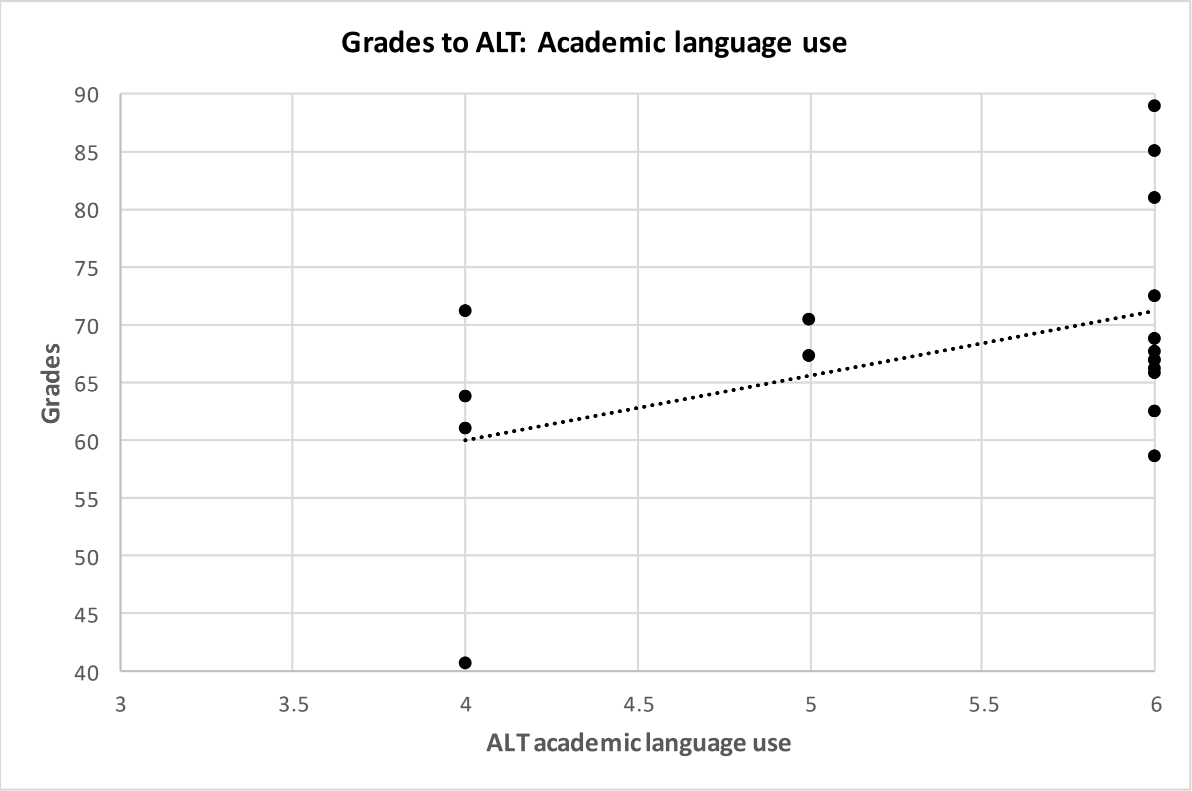

Figure 5: Grades to CAEALT, Academic language use

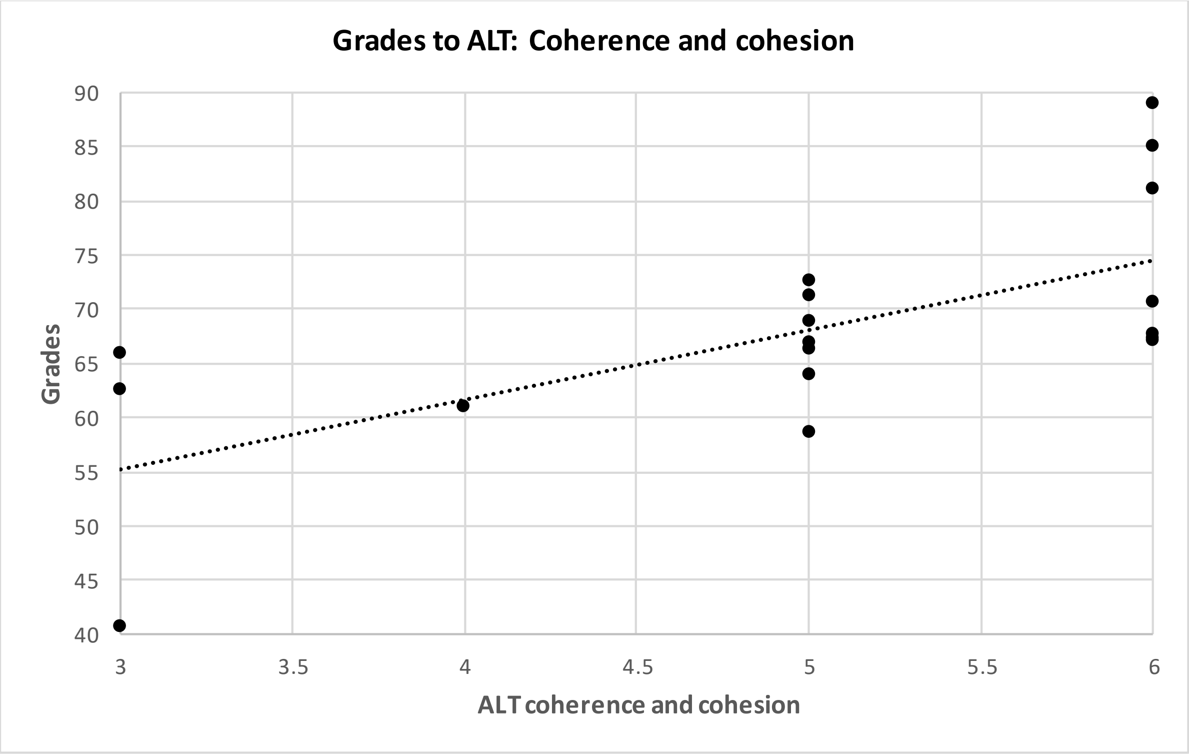

Figure 6: Grades to CAEALT, coherence and cohesion

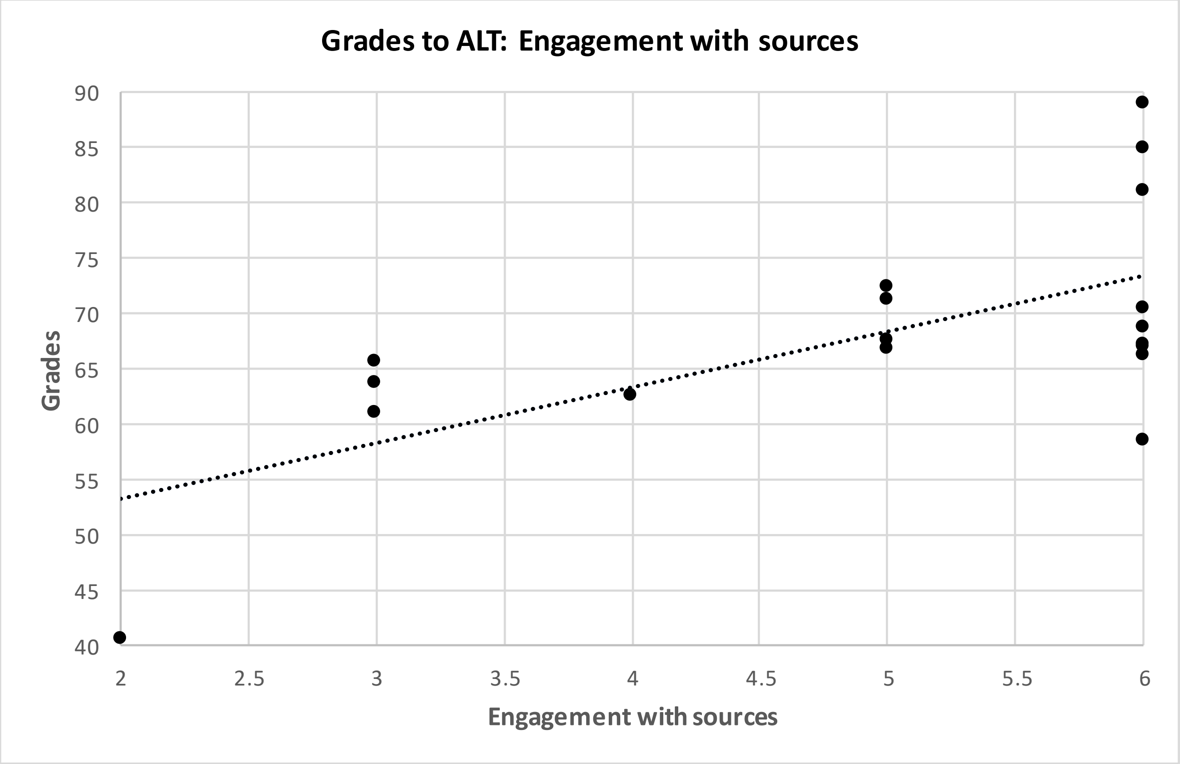

Figure 7: Grades to CAEALT, engagement with sources

Figures 3-7 show the correlation between university grades and CAEALT results (overall and by each marking criterion). As discussed in section 3.2.5.1, a correlation coefficient of over 0.2 is considered acceptable when analysing admissions tests. This study will follow the conventional thresholds of a p-value of 0.05 and effect size of 0.50.

	Correlation coefficient (Kendall's Tau, τ)	P-value	Effect size (Cohen's d)
Overall	0.518	0.04	8.42
Argument	0.344	0.074	8.4
Coherence & cohesion	0.576	0.003	8.43
Academic language use	0.235	0.233	8.39
Engagement with sources	0.467	0.014	8.42

Table 17: Grades to CAEALT results

Three of the five measures show a significant correlation ($p < 0.05$) with university grades: the overall CAEALT mark, coherence and cohesion, engagement with sources. Argument has a p-value slightly higher than usually accepted, but the correlation coefficient and effect size suggest that this category should be tentatively included as approaching significance. Academic language use is the weakest of these measures and the correlation with the grades here is not significant: a larger study is necessary to determine if this is a possible measure of academic literacy.

5.1.1.4 Self- and tutor assessment to ALT

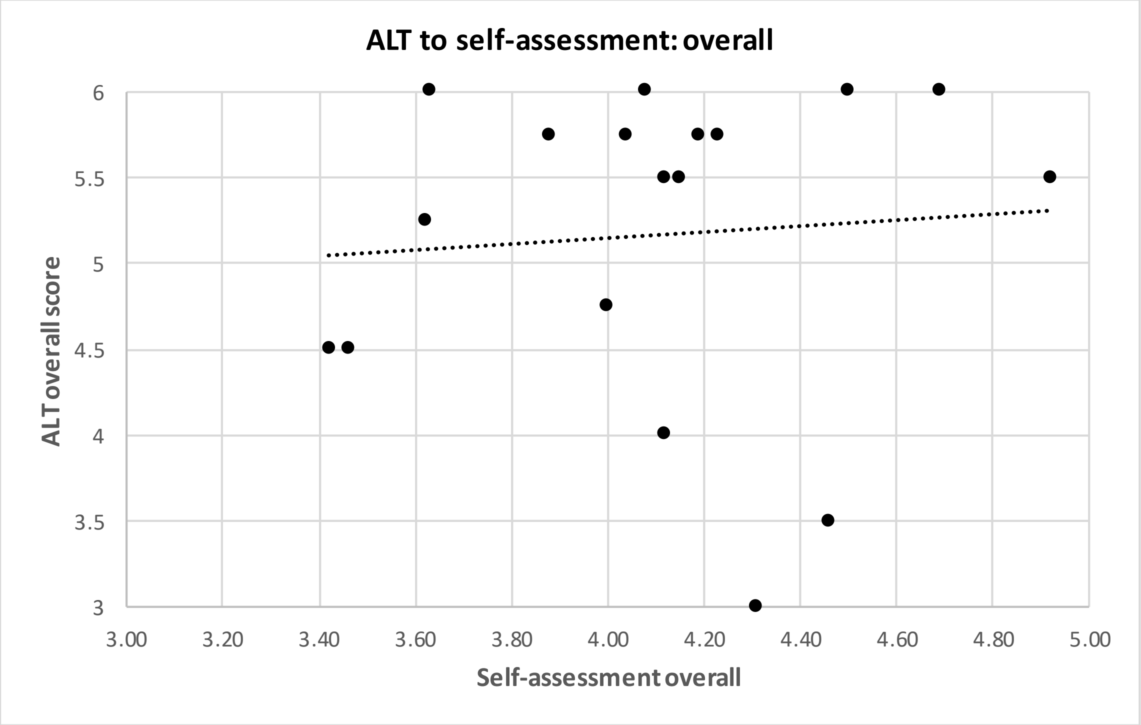

Figure 8: Self-assessment to CAEALT, overall

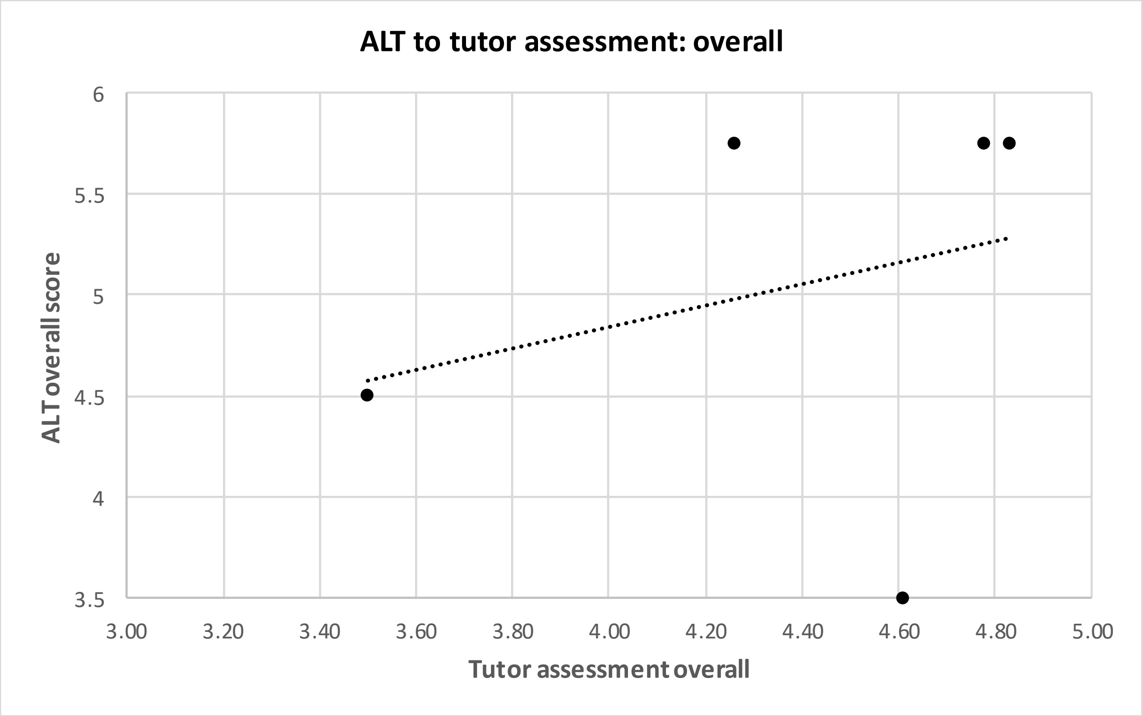

Figure 9: tutor assessment to CAEALT, overall

	Self-assessment (n=18)			Tutor assessment (n=5)		
	Correlation coefficient (τ)	P-value	Effect size (Cohen’s d)	Correlation coefficient (τ)	P-value	Effect size (Cohen’s d)
Overall	0.125	0.488	1.51	0.359	0.405	1.02
Argument	0.042	0.827	1.6	-0.096	0.810	0.78
Coherence & cohesion	-0.124	0.515	1.12	-0.118	0.788	0.7
Academic language use	0.108	0.582	1.91	0.545	0.150	1.45
Engagement with sources	-0.187	0.324	0.77	0.320	0.412	0.54

Table 18: CAEALT results to self- and tutor assessment

Figure 8 and Figure 9 show the relationship between CAEALT results and overall self-/tutor assessment. Correlation coefficients and p-values overall and for each criterion are in Table 18. Neither self- nor tutor assessment showed any significant correlation with the CAEALT scores. Further discussion of this lack of correlation will take place in the discussion section; as previously stated, the number of tutor assessments available is extremely small (n = 5) and therefore no meaningful conclusions can be drawn from this data.

Correlations were also calculated between self-assessment and university grades and between tutor assessment and grades, both for the overall results and for each criterion. However, no further correlations of interest were found.

5.1.2 Qualitative data

As previously stated, qualitative data came from the test-taking experience questionnaire, from additional comments from students and tutors, and from candidate scripts. The last of these will be discussed in the case studies forming section 5.2.1. Very few student or tutor comments were submitted as part of the self-assessment. However, the test-taking experience questionnaire was more fruitful. Comments from this questionnaire were coded into several categories: all categories and comments can be found in appendix 8.9. I will comment on three key categories below.

Selection and critical engagement were only included under similarities; that is, no differences were mentioned. Candidates tended to comment on either selection or critical engagement – only one commented on both, suggesting that they may be seen as mutually exclusive. Critical engagement is unique in that every comment used the words 'evaluate' or 'evaluation', with little glossing of exactly how this term was being interpreted.

Comments on time pressure fell into two categories:

- Four participants commented on the time available under exam conditions, with two participants identifying the given time as similar, and one saying it was more generous than her university exams. One noted she never read under exam conditions and 'was not trained to read and quickly understand material'
- Seven participants noted they would usually write essays over a multiple-day period, with that longer time used to 'find an angle I'd like to explore', 'find resources', 'consider sources', arguments', 'clarify / amplify [sources] ideas', 'plan' and 'proof-read'

Comments on source material focused around the style, density and complexity of the writing, the quantity of text and the number of sources. The most common difference was that sources were already selected, rather than the participant selecting them, sometimes from a reading list, for themselves.

5.2 RQ2: Discussion and case studies

Overall, the quantitative results suggest that there is a case for a concurrently valid correlation between the CAEALT and university grades, for a combined sample of undergraduate and postgraduate students, with an overall correlation coefficient of 0.518 (bearing in mind the suggested coefficients for predictive validity tests in section 3.2.5.1). Conclusions from the quantitative results must be tentative due to the small sample size and limited range of university grades, but the indications are promising and suggest that more extensive validation of this test may reach the same conclusion.

Two strong correlations were found between university grades and CAEALT marking criteria: coherence and cohesion (τ =0.576) and engagement with sources (τ=0.467). There is a likelihood of some correlation in the area of argument (τ=0.344, p = 0.074), but no correlation was found for academic language use (τ=0.235, p=0.233). As only a small sample size was collected, these conclusions from the quantitative results will be discussed in the context of qualitative case studies. Following these, the main areas in each marking category will be summarised and possible interpretations given.

5.2.1 Case studies

To gain further insights into the relationship between the CAEALT construct and the other measures of academic ability, I looked at four candidates in greater detail: two with strong CAEALT scores, and two with low CAEALT scores. Candidates 7 and 10 received the lowest scores in the sample; candidate 14 was one of four candidates who received maximum marks in the CAEALT and, in my opinion, the strongest of those four (although no marking criteria are available at this level, making this judgement relatively subjective). Candidate 5 was one of the joint strongest candidates for which tutor feedback was available.

As the purpose of these case studies is to contribute to RQ2, subskills were selected for closer examination based on either substantial differences between the different measures of ability available, such as a high tutor rating for a particular subskill but a low CAEALT mark in that category, or based on areas of similarity between those ratings. Each case study takes the structure of first discussing the subskills that draw a picture of the candidate's academic literacy, then examining those subskills that may provide an insight into gaps in the CAEALT construct.

	Candidate number	Overall CAEALT result	Argument	Coherence & Cohesion	Academic language use	Engagement with sources
Low	7	3.5	3	4	4	3
Low	10	3	3	3	4	2
High	5	5.75	6	5	6	6
High	14	6	6	6	6	6

Table 19: CAEALT marks for case-study candidates

5.2.2 Standardisation of measurements

Before analysis of individual case studies, a brief discussion of the measures used is required. The three measures (CAEALT scores, self-assessment, tutor assessment) face the fundamental issue of there being no standardisation between either the self- and tutor assessments for an individual candidate, or between two self- (or tutor) assessments. That is, one cannot assume a 5 in one is equivalent to a 5 in another. Therefore, the only meaningful comparison that can be made between measures for each candidate, or across candidates, is relative rather than absolute: a score can be described as either above or below average, or as being of greater or lesser magnitude.

It could be argued that, as scores were reported against a scale from strongly disagree to strongly agree, this provided some measure of standardisation and therefore some comparison is possible. However, in my opinion, the standards informing these decisions, i.e. what two people consider competence in a skill, is subject to significant variation and therefore cannot unequivocally be used at the basis for comparison.

Because the focus is on relative rather than absolute scores, the self- and tutor assessment scores are given first as raw scores, but then have been normalised by subtracting each subskill score from the overall mean for either the individual student or the tutor (depending on whether the self- or tutor assessment is in question), and dividing by the overall standard deviation; this gives us a score where zero represents the average across all subskills, a positive score an above average score, a negative score below average, and the magnitude the difference from the average.

For example, a candidate may receive the following scores:

	Self-assessment	Tutor assessment
1.1.1 Generate ideas for writing by using texts in addition to past experience or observations	3 (-0.37)	3 (0.10)
2.3.2 Make connections to related topics, information or prior knowledge, even when they are not obvious.	5 (0.22)	4 (0.34)

For the self-assessment scores, the normalised figures indicate that the candidate rated themselves below their self-assessment average for subskill 1.1.1, and above average for subskill 2.3.2. This can be compared with the tutor assessment, where both scores are above this tutor's assessment average. Naturally, the scale on which magnitude is represented does not equate to the self- or tutor assessment scale: a normalised subskill score of 1 when a candidate has an average score of 4 does not indicate that this subskill scored a 5.

As mentioned above, raw scores for each subskill have also been included. As there is in fact little variation between the scores given to different subskills, this provides a useful check as to the actual magnitude of the differences under discussion. However, please note the caution below on comparison of raw scores.

The CAEALT scores have not been normalised as, since very few numbers are concerned, the necessary scores can be intuitively calculated by the reader. The overall average is given by each subskill to facilitate this.

When we come to comparison between candidates, CAEALT scores can easily be compared as they were marked by the same examiner according to the same scale. This is not the case when it comes to self- and tutor assessment raw scores: one should not assume equivalence between a score of 5 given by one tutor to Candidate A and a score of 5 given by another to Candidate B. When it comes to relative (normalised) scores, some comparison is possible: if two candidates give themselves a normalised score of 0.5 for a particular subskill, they both consider themselves to be better by the same amount from their average. However, there is substantial room for variation here and so any such comparison must be tentative: a 0.7 for one candidate and a 0.5 for another could indicate that they are of the same absolute ability for this skill, but that the former has a lower average score overall, or that they are of the same absolute ability overall and that the first candidate is stronger at this skill.

Please note that when the average scores for each marking category were calculated, the scores were averaged first and then normalised.

5.2.3 Lower-scoring candidates

5.2.3.1 Candidate 7

Candidate 7 is 19 years old, female and from Hungary. With Hungarian as a first language, she also speaks English and Italian. She entered university with an IELTS qualification, scoring 7.5 overall, 9 in Reading and 7 in Writing. Her tutor notes that her ‘first language is not English but she copes very well’.

She is in the 2nd year of an undergraduate degree in Psychological and Behavioural Sciences at the University of Cambridge; in her exams at the end of her first year she scored between 60 and 62, with her average mark as 61 (a low 2.1). Her previous highest qualifications were A-level equivalent.

The candidate's essay script forms appendix 8.10.1.

Candidate 7 scored the second lowest mark in the CAEALT and received the third lowest grades among the students sampled. Across all categories of the CAEALT, candidate 7 scored lower than the average. Her joint strongest and joint weakest skills are the same as for the candidate sample as a whole: academic language use and engagement with sources respectively, suggesting that her skills profile is typical of the sample. However, both her self- and her tutor assessment were consistently above average in all categories but one, that below-average category being different between the two assessments (argument in the case of the self-assessment, engagement with sources in the case of the tutor assessment).

There are several areas which have been identified by at least two of the three measurements (CAEALT mark, self-assessment, tutor assessment) as weaker. These are primarily in the categories of argument and engagement with sources.

Argument

CAEALT	Self assess.	Tutor assess.	

Argument: 3 (overall average: 3.5)	4 (-0.71)	4 (-0.97)	6.1.1 Develop thesis convincingly with well-chosen examples, reasons and logic

For argument, while the self- and tutor assessment are still strong scores, they are slightly below average, and so all three measures are in agreement in identifying this as weaker skill.

Subskill 6.1.1 is particularly problematic in the candidate script, especially as the candidate's thesis is implicit only, and does not follow on from the body of the essay. However, the key issue relates to the quantity and quality of implied information, as exemplified in paragraph 7.

> [The findings of] Oimiss-Penuela, Benneworth and Castro-Martinez (2015) …also rather suggest a revival [sic] of validity guidelines. They proposed that humanities research doesn't need as much external validity as sciences do, since applicability of the results is relatively smaller. In addition, they cite Cassity and Aug (2006) who wrote that humanities are less related to business innovation, and the authors also claimed that there is less demand for humanities research than for science research.
>
> Candidate 7, paragraph 7

In this example, she combines two arguments: the first is that humanities does not need as much external validity as the sciences, because the applicability of the results is smaller. The second is that, because the need for external validity is smaller, validity guidelines should be revised. There are two main issues relating to subskill 1.4 in this passage. The first issue is that there is a substantial amount of implied and unstated information which is not included either in the candidate's essay or in the source texts: the terms 'external validity' and 'validity guidelines' are never defined and are not intuitively understood by the reader; in the second argument it is not stated in what way the validity guidelines should be revised. It is also not clear how the final two examples (business innovation, less demand) relate to either of the arguments.

However, the bigger issue relates to the first argument, whose structure is as follows (It is worth noting here that this is not an argument that exists in the source texts):

- Premise: humanities research is less widely applicable
- Inference (implied): Things that are not used as widely do not need external validity
- Conclusion: humanities does not need as much external validity

The lack of explicit inference is problematic as it is not easy to deduce from the given information. The bigger issue is that this implied inference is a controversial one which requires more defence than is given: a reader may argue that there should be no such relationship between use and quality because, for example, quantity of use has no relationship to the importance of use: a paper commissioned by a government department may be read by less than 100 people, but will likely be more influential than a newspaper article read by millions. By missing out key information, candidate 7 has made her arguments harder to understand and has not provided any evidence for the most controversial part of her first argument.

The candidate scored very differently in her self-/tutor assessment (strong) and the CAEALT (weak). It seems unlikely that this is a manifestation of differences in genre and argumentation between faculties: when asked how the CAEALT was similar to activities in her everyday university life, the candidate wrote, ‘having to express my own opinion on a certain topic with using evidence from other resources’ and ‘having to write in an argumentative style’. This suggests that the expected structure of an argument is not a factor. However, it is possible that the type of issue to be defended is more abstract in the CAEALT, as she wrote that ‘I rather [sic] read experimental papers’, and ‘I write essays on topics that are more relevant to my subject’.

Engagement with sources

CAEALT	Self assess.	Tutor assess.	
Engagement with sources: 3	4 (-0.71)	4 (-0.97)	2.3.2 Make connections to related topics, information or prior knowledge, even when they are not obvious.
	4 (-0.71)	4 (-0.97)	2.3.5 Understand separate ideas and then be able to see how these ideas form a whole

2.3.2 and 2.3.5 are concerned with idea generation through synthesis; that is, with comparing and contrasting different parts of a source text, or different source texts, to uncover new ideas not directly contained in those sources. These subskills are demonstrated to a small extent in the candidate’s script: for the most part the ideas in the sources are simply described or listed, and the limited number of original ideas presented are mostly extensions of one source idea rather than generated through synthesis.

There is one example of limited synthesis: in paragraph 2 she combines an idea from a source text with the essay question to create the new idea (underlined).

> Small (2013)… argues that research in the field …is measurable in terms of the income produced by bookshops, museums, heritage sites, theatres etc. Therefore, applying the same evaluative criteria as before is useful so that the economic impact can be distinguished between science and humanities research.
>
> Candidate 7: paragraph 2

Candidate 7 takes the idea from the essay question of 'using the same criteria' for humanities and sciences and uses this as a frame through which to view source text 1, '[the humanities] make a significant contribution to the knowledge economy and to the economy proper – measurable in terms of the benefits to GDP, footfalls in bookshops, museums, theatres, heritage sites, and so forth', thus drawing out the concept that the quantitative measure of benefit to GDP could be used to measure change in evaluative criteria. However, this is the only example of true synthesis in the essay.

In her test-taking experience questionnaire, she commented that a similarity between the CAEALT and her everyday university experience was 'evaluation of multiple resources' and 'having to express my own opinion on a certain topic with using evidence from other resources'. In her CAEALT script she has demonstrated the second of these, but not the first. It would be interesting to discuss with her her precise understanding of 'evaluation', to see if her definition was in line with that given in this thesis.

It could be suggested that synthesising ideas from multiple sources is a difficult area to elicit under exam conditions, particularly when a candidate is faced with a significant reading load and an essay task in a subject that may not be related to their own.

CAEALT	Self assess.	Tutor assess.	
Engagement with sources: 3	5 (0.83)	4 (-0.97)	4.1.1 Use of quotation, paraphrase and summaries to avoid plagiarism

Inappropriate quotation, paraphrase and summary (subskill 4.1.1) is a key weakness of the script. The CAEALT instructs the candidates to reference sources and provide in-text citations, but Candidate 7 often directly lifts from the source material or makes minimal changes, without acknowledging these direct quotations as such (of the 902-word script, 402 words are quotations, paraphrases/summaries or in-text referencing). Additionally, without a direct comparison with the original sources there is often little to indicate which ideas are taken from the sources and which are original.

In the candidate’s script extract below, paraphrases of the source materials are in bold and direct lifting or near-lifting underlined.

Candidate 7 script	Source text(s)
The bibliometric measures currently employed are used in practice **for evaluating the validity and social impact of papers**, and evaluating this into, for example, bases of university funding systems in several countries. [paragraph 1]	Bibliometric indicators are used to compare and evaluate research performance in the life sciences and natural sciences… and are employed in the performance-based university research funding systems of several countries. [Hug, Ochsner and Daniel, 2014, paragraph 1]
The second argument against using the same criteria is that since citation counts are also	Hose (2009, p. 95), a scholar of Greek philology, argues that citation counts ‘have

included, there is a tendency to favour spectacular research and neglect ones from more marginalised fields. Another problem supporting this argument is the fact that authors often use self-citation or cite friends exclusively and this manipulate reliability (Charle, 2009). [Paragraph 5]	the tendency to favour spectacular (and given certain circumstances, erroneous) results, and penalize fundamental research and sustainable results as well as those doing research in marginal fields' (own translation). Moreover, Charle (2009) claims that citation counts can easily be manipulated by self-citations or by citing friends excessively." [Hug, Ochsner and Daniel, 2014, paragraph 4]

This subskill, rated highly by the student, gives an insight into what she considers to be appropriate source use. As it seems unlikely that she would consider unattributed use of source materials appropriate, and that she can have reached this stage of study without being corrected on this, it is possible that this is an artefact of the scripts being presented as part of an exam booklet – that she felt she was expected to draw on the material in this way. The CAEALT explicitly requires referencing, but does not mention specifics such as the attribution of direct quotations, and it is possible that she assumed they were not necessary. Comparisons would need to be made to other written work to establish this.

5.2.3.2 Possible mismatches between CAEALT construct and AL taxonomy

1.1.2 Fully understand essay questions	
CAEALT mark	3
Self-assessment	4 (-0.71)
Tutor assessment	5 (0.51)

Candidate 7 received slightly different self- and tutor ratings. A possible reason for this could be based on different interpretations of the subskill's wording: I suggest that a student in an L2 context may interpret it in terms of understanding the grammar and lexis making up the question (especially as they were also asked for IELTS scores as part of their demographic information); a tutor may be more inclined to interpret this as understanding the task requirements/genre/readership (macro-planning). A third interpretation is possible: the CAEALT marking criteria require the presentation of a clear position, with little to no irrelevant content. Another possible explanation for the difference in ratings may stem from the difference between process and product: the student will experience difficulties in the process of producing an essay that are not reflected in the final product seen by the tutor.

Whichever explanation is the case, it seems that this is not a subskill that can substantially differentiate between candidates beyond the binary of understood / didn't understand the question (an issue that will be returned to below). The CAEALT marking criteria mentioned above are more fully and explicitly covered under subskills 2.3.2 (non-obvious connections to related information), 2.3.5 (how ideas form a whole) and 6.1.1 (develop thesis convincingly).

CAEALT	Self assess.	Tutor assess.	
Engagement with sources: 3	5 (0.83)	4 (-0.97)	2.1.5. I can always identify the attitude/opinion of the author
	5 (0.83)	3 (-2.45)	2.2.1 I can always identify the main thesis of a whole text
	4 (-0.71)	4 (-0.97)	2.2.2. I can always identify the major and subordinate ideas in a particular passage of text

The issue of how to appropriately prove understanding of input material (as in the section above) also applies to the subskills surrounding comprehension of input material. In general, the candidate's script does not show issues with comprehension that reflect the tutor's below average ratings of these subskills. There are instances of misrepresentation of source texts, presumably due to a lack of comprehension (discussed above), which fall under subskill 2.2.2 (major and subordinate ideas – identified by both candidate and tutor as a weaker subskill), but nothing to reflect the particularly poor rating for subskill 2.2.1 (identifying main thesis). This suggests that, as with subskill 1.1.2 above, the CAEALT does not substantially differentiate between candidates in this subskill. This is reflected in the marking criteria, which have little focus on comprehension of sources beyond a mention of possible misrepresentation.

***

Overall, the key areas of the candidate's academic literacy that have impacted on the CAEALT mark are in the categories of a clearly-structured argument and of appropriate quotation and paraphrase. Subskill 6.1.1 (develop thesis convincingly) was identified by the CAEALT, self-assessment and tutor assessment as weak.

Generation of new ideas through synthesis (subskills 2.3.2, 2.3.5) was limited. While this was demonstrated more fully by other candidates (see section 5.2.4.1), it could be that it is more common when a student performs their own literature search as they are actively seeking answers to self-generated questions.

Subskill 4.1.1 (quotation, paraphrase) was a weakness of the CAEALT script which was not reflected in the student or tutor ratings. I have suggested that this is an artefact of the exam format – of sources being provided – rather than indicating a lack of knowledge of appropriate source use, but an interview would be necessary to establish if this is the case.

While the tutor feedback suggests that comprehension of sources may be an issue, this has not been apparent in the candidate's script, suggesting that this may not discriminate in the CAEALT.

Candidate 7's IELTS scores should be briefly mentioned here: On university entry two years ago she scored 9 in Reading (the highest band score) and 7 for Writing, with a 7.5 overall score (the higher end of C1), meaning that she equalled the typical minimu m LP requirements for her university. However, as the CAEALT is aimed at C1-C2, her IELTS scores can be considered at the lower end of the range tested. It should also be noted that, while LP forms part of the CAEALT construct, it is only a part. The writing to sources requirement makes the task expectation quite different, becoming more reliant on evidence than exhortation, and there is a greater emphasis on argument. For these reasons, her IELTS score is not incompatible with her CAEALT scores.

For candidate 10, there are several areas where the CAEALT marks and the self- / tutor assessment are in agreement. For subskills 6.1.1 (develop thesis convincingly), 2.3.2 (connect related topics) and 2.3.5 (coherent thesis from separate ideas), the self- and tutor assessments are – while not as low as the CAEALT marks – below the average for this candidate. This suggests that the candidate performance elicited by the CAEALT is representative of the real-life construct in these areas. For the remaining subskills examined in this case study, I have suggested some reasons why the self- and tutor assessment and the CAEALT may not agree. For 2.1.5 (identify attitude of author), 2.2.1 (identify main thesis) and 2.2.2 (major and subordinate ideas), it seems likely that the skill may not substantially differentiate between candidates. For subskill 4.1.1 (quote, paraphrase, summarise) I suggest that the source use elicited in the CAEALT is unlikely to be representative of her real-life use. For these subskills, then, an alternative means of assessment needs to be used.

5.2.3.3 Candidate 10

Candidate 10 is 22 years old, female and from Poland. Polish is her first language; she also speaks English, German and Spanish. She is in the final year of a degree in Criminology at Anglia Ruskin University. Her university marks are the lowest in the sample, with an average of 40.61 (she received a mark of 40 in all courses except one – this one course is not, on the surface, different from the other courses).

No tutor feedback is available for this candidate for reasons discussed in section 3.2.4. This candidate has not taken any LP exams. Her previous highest qualifications were A-level equivalent.

She notes that the CAEALT was very similar to the activities in her everyday university life: ‘most of the modules on my Criminology course have had an essay approach to assessment’, but also that ‘I have only completed two exams throughout the three years.’ This suggests that she has written the majority of her essays at home, and therefore that her score in the CAEALT may be less representative of her everyday university performance; she has not often (if at all) been required to write essays under timed conditions.

Candidate 10 received both the lowest CAEALT mark and the lowest university grade in the sample. While comparisons between candidates using self-assessment scores should be treated with great caution for the reasons outlined earlier, it is interesting to note that she has the 5th strongest self-assessment score across both undergraduates and postgraduates, suggesting that there may be a gap between her expectations of appropriate performance and that of others. In all categories of the CAEALT, candidate 10 scored lower than average. Both her strongest and weakest skills are the same as for the sample as a whole: academic language use and engagement with sources respectively.

Her script forms appendix 8.10.2.

Engagement with sources

CAEALT	Self assess.	
Engagement with sources: 2	5	2.3.8 Synthesize information from several sources and incorporate it into a writing assignment
	4	4.1.1 Use of quotation, paraphrase and summaries to avoid plagiarism

As with candidate 7, the key weakness of the script in this category is significant overuse of direct quotation (subskill 4.1.1). Of the 611 words of the script, 282 are direct quotations (and one paraphrase) – that is, 46% of the text is directly lifted from the input, of which 128 words are unattributed quotations, and 118 appear in the region of an attribution, but it is not made clear that they are direct quotations as there is no use of quotation marks. Paragraph 3 below exemplifies this.

> The value of humanities has been examined by (Small, 2013). There are five claims established. The first is that the value of humanities is meaningful since they study the meaning-making practices of the culture. Secondly, there is a significant pressure on how governments commonly understand use and prioritize the scale of economic usefulness. (Small, 2013) Thirdly, (Small, 2013) takes stance that the humanities have a contribution to make to our general happiness. Furthermore, the fourth claim 'democracy needs us' is the most ambitions argument now regularly heard for the humanities in Britain. The final claim is that the humanities matter for their own sake. (Small, 2013) The five arguments have been influential in ancient history and maintain persuasive power. It is an easy task to evaluate the work of (Small, 2013), since the scholar's publication is of significantly large content, in comparison to (Olmos and Penuels, 2013).

Candidate 10, paragraph 3 (direct quotations are underlined)

Interestingly, the text does show skill in combining these chunks of text in a relatively coherent way (subskill 4.8, combine ideas from several sources); this is a skill that the candidate scored herself highly on.

Possible reasons for the difference in rating between the self-/tutor assessment and the CAEALT mark are discussed under candidate 7 – as with candidate 7, it seems unlikely that a candidate can have reached the end of her course without being aware of referencing conventions. As this candidate has lower scores in her university marks it is more possible that this inappropriate source use could take place in real-life: however, if that were so the student would be aware of this and assess herself accordingly. I suggest, therefore, that the fact this issue occurs in two candidate scripts suggests that this is an issue with the test format.

Coherence and cohesion

CAEALT	Self assess.	
Coherence and cohesion: 3	3	4.2.1 Develop main point or thesis

Candidate 10 rated herself joint lowest in subskill 4.2.1 (develop main point): this subskill targets the ability of the candidate to provide further detail to describe and flesh out an idea. The candidate's lower opinion of her ability is backed up by analysis of her script: she has a tendency to group together loosely-related points and facts rather than explore any individual point or fact in detail. Thus, this subskill is connected to the two subskills discussed previously (2.3.8 – combining ideas, 4.1.1 – appropriate quotation) in partly being an issue of source use: As with direct quotation, the candidate appears to think that simply relaying the information contained in the source material is sufficient (knowledge telling) rather than understanding the necessity of synthesising the ideas from the source texts to form a coherent new text (knowledge transforming). The other aspect of this subskill, that is not related to source use, is the ability to express and develop own ideas; unfortunately as the script is highly reliant on concepts from the sources, this other aspect is not demonstrated in this instance.

***

Overall, the two granular measures of candidate 10's academic literacy (self-assessment and CAEALT) were not often in agreement. Of those selected for analysis here, significant overuse of direct quotation (4.1.1) was the key weakness of the CAEALT script, and I have hypothesised that this is a consequence of the exam context.

However, this analysis must be seen in the light of the candidate's university grades, which were in line with her CAEALT scores: in both cases she was the weakest in this sample. It is possible that either her expectations of acceptable performance were significantly lower than those of the CAEALT, or that she chose to report higher than her actual perceived ability.

5.2.3.4 Possible mismatches between CAEALT construct and AL taxonomy

CAEALT	Self assess.	
Academic language use: 4	5 (1.12)	5.2.1 Proofread to eliminate errors in grammar, mechanics and spelling, using standard English conventions

Subskill 5.2.1 relates to revising rather than monitoring, i.e. revisiting the text after writing is complete rather than checking the text during the act of writing. Two revisions have been made in candidate 10's script, both in fact creating errors rather than correcting them:

'the ~~role~~ studies of arts and humanities is questioned by…'

'The content, as observed ~~in~~ between publications of [Small and Olmos-Penuela et al]'

In the case of this script, subskill 5.2.1 is not displayed to advantage. However, this does not have a significant effect on the CAEALT mark as this subskill is not explicitly tested; the text is accurate overall and is therefore not penalised under the CAEALT markscheme. It should also be noted that, due to the large proportion of lifted text, formulaic language and direct quotation, this small number of errors becomes a rather larger proportion of the original text: of the eleven significant-size chunks of text remaining (of five words or over), six have an error or an ambiguity relating to word choice.

5.2.4 Higher-scoring students

As previously mentioned, four candidates received the maximum marks in the CAEALT, suggesting that the ceiling effect (where normal distribution is distorted by a maximum mark) is a factor in a lower correlation. In fact, fifteen out of the eighteen candidates received a mark of 6 in at least one category (the most common category being academic language use), suggesting that the ceiling effect could have a significant effect on the correlations recorded.

The discussion below takes place in the context of highly-performing students at a higher level than the CAEALT is aimed. To understand whether the CAEALT may be able to elicit the skills necessary at the end of a Master’s degree or the beginning of a PhD – that is, the full range of academic skills – I will include in the discussion those skills which are beyond the top end of the CAEALT markscheme. Therefore, while the analysis below will pick out areas where the full range of academic skills is not elicited, the reader should understand that postgraduate entry skills were fully demonstrated in the scripts.

5.2.4.1 Candidate 14

Candidate 14 is female, 39 years old and Greek. She also speaks English, French, Spanish, Italian, Portuguese, Russian, Catalan, Japanese and German. She has just finished an MA in Translation Theory at Ionian University, Corfu; her undergraduate degree was in Spanish Language. She took Cambridge Proficiency in 1996, receiving a B. Her postgraduate marks were out of 10, and so have been translated into out of 100 for the purposes of the quantitative analysis. Out of the modules she took, Spanish language was her strongest, (10/10) and Latin American Literature her weakest (7/10). Her average mark was 8.9/10. Candidate 14 currently works as a proofreader and editor. As with candidate 10, no tutor feedback is available. Her script forms appendix 8.10.3.

In all categories of the CAEALT, candidate 14 scored above average. However, she has the fourth lowest self-assessment score across the whole sample and the second lowest postgraduate self-assessment rating. It is interesting, but not surprising, to notice that the subskills that she uses regularly in her working life, and thus those on which she is receiving regular feedback that she meets the standards expected of her, are those she has rated herself at a 5: these are three categories in academic language use, namely, 1.1.3 consider audience and purpose, 2.1.4 decipher the meaning of vocabulary and 5.2.1 proofread to eliminate errors. Other categories are rated lower. This suggests that there is a significant gap between her expectations and those expected of CAEALT candidates – that she expects higher of herself than is required in the CAEALT. As someone who has recently finished a Master's degree, this is congruent as she is working to the standards expected of her at this level, rather than the postgraduate entry skills the CAEALT is targeted at.

In terms of weaker subskills, the clearest picture of candidate 14's self-assessment scores emerges when her self-assessment scores are sorted by cognitive process according to Chan's 2013 reading-into-writing framework: she has consistently rated herself poorly on subskills related to meaning and discourse construction: connecting and generating (see section 2.1.3); the two which are relevant to this case study are given below.

Synthesising ideas in one source

CAEALT	Self assess.	
Engagement with sources: 6	3 (-0.71)	2.3.5 Understand separate ideas and then be able to see how these ideas form a whole
	2 (-1.86)	2.3.8 Synthesize information from several sources and incorporate it into a writing assignment

Listing of ideas taken from the sources, without transformation, based around a theme is one of the key methods of synthesis in the candidate's script. This is exemplified in the text extract below, which is a synthesis of some key ideas from source C. Source C is made up of a series of tables, each table presenting a list of one-sentence facts supporting a different point of view. The first table presents possible reasons why humanities are less valuable than sciences, and the second table possible reasons why they are differently valuable. The third table is also concerned with the relative value of the humanities and sciences, but is concerned with how such differences could be operationalised to allow assessment of each i.e. possible real-life consequences of the differences. Candidate 14 has selected from this list of possible facts to produce the extract below. The numbers in brackets represent which of the three tables the idea was taken from, with the letters separating different facts within each table.

Candidate 14, script extract	Ideas from source
… while Humanities research relates to smaller scales when compared to sciences research [1]… there is value in promoting the former and strive for its fair evaluation. This smaller scale to which humanities research usually relates also means that the profile of the Humanities research users is very different from the profile of science research users [2,3]. Humanities researchers work more directly with a broad range of users [2a, 2b, 2c, 3c], who come mainly from the public [2b, 3b], and voluntary sectors [3c] and, more often than not, this is limited to a national level [3a], while science researchers work mainly with firms [2c] and more often on an international level [3a] (Olmos-Penuela et al, 2015).	1. 'Humanities research is less scalable with less applicability to other contexts' 2a. 'Humanities researchers work directly with users, but often in ways that are less visible and formalised' 2b. 'Humanities researchers communicate with publics via commentary, whilst publics are interested in the business of science' 2c. 'Humanities researchers tend to work with a much broader range of users than scientists who mainly work with firms' 3a. 'The rate of involvement with national users compared to international users is higher for humanities researchers than for science researchers' 3b. Humanities researchers spend more time in popularisation activities than science researchers 3c. 'Humanities researchers collaborate less than scientists with firms and more with public and voluntary sectors'

In this extract the candidate has picked out one of the thematic links between the tables: that of the differences in scale and audience for humanities and science research, and successfully paraphrased and summarised the key facts in a few coherent sentences. At this level – that of connecting ideas – her synthesis is successful. Her final sentence in particular is a good example of recasting of the source information: she takes several facts that are presented separately in the source material and rearranges them around a sentence framework directly comparing the humanities and sciences.

Synthesising ideas across multiple sources

Candidate 14's CAEALT script does not demonstrate any synthesis of ideas across sources (as above, please notice that this is not required for full CAEALT marks. Whether this should be the case will be discussed in section 6): Each of her arguments has been taken directly from only one of the sources. For example, her second argument – that the humanities and the sciences cannot be assessed using the same criteria – is entirely taken from source 2, with no commentary or contrast with information elsewhere in the sources.

Candidate 14 extract 2	Source 2
The second reason why humanities research cannot be evaluated using the same criteria as the ones used to evaluate science research lies in the methods that have been put forward so far. Most of these methods have been borrowed from the natural sciences (Hug et al, 2014), which renders them unsuitable. This is due to the non-linear fashion in which humanities research progresses and also the more evident fact that a lot of humanities research cannot be easily quantified. What scholars stress is that the part of humanities research that actually is measurable, is not usually significant and that indicators typically used to quantify research impact provide little new information to the assessor.	Bibliometric indicators are not well-suited to determine the quantity and quality of humanities' research or to assess it… Vec (2009), a legal scholar, claims that 'a lot of evaluation systems were modelled after the natural sciences'… Lack (2008), a literature scholar, asserts that existing evaluation procedures and indicators are based on a natural sciences' linear understanding of progress and, therefore, asks for tools that can cope with the humanities' conception of increasing knowledge… [Academics Australia have] widespread reservations regarding the quantification of research quality in the humanities… Other humanities scholars do not deny that research quality or performance can be expressed quantitatively, but point out that measurable output is not important in the humanities and indicators convey information that is already widely known.

In this argument, candidate 14 leaves out a key aspect that is in both sources 1 and 3: that one of the key reasons that methods for evaluating research in the sciences do not suit the humanities is that the latter has a 'distinctive understanding of what constitutes knowledge – differentiating them from the social sciences and the sciences where the emphasis on subjectivity is less strong' (source 1) and that 'there are no 'right' answers to humanities questions, just opinions' (source 3).

In defence of the candidate's lack of synthesis, the sources are quite different in their messages: each deal with a different aspect of the contrast between the arts and the humanities (source 1 with arguments for the value of the humanities, source 2 with the use of bibliometric indicators for the humanities, source 3 as above) and so there are no disagreements to be resolved between the main theses, and few within the bodies of the texts.

There is one disagreement between the given sources: this in fact has been transferred to the candidate script without the disagreement being noticed or resolved. In paragraph 2 (taken from source 1), the script reads:

> 'Advocates of the value of the humanities and their impact on societies have argued that the benefits from humanities research can be translated ... into measurable goods, such as increase in GDP'
>
> Candidate 14 extract 3

While in paragraph 3 (taken from source 2):

> What scholars stress is that the part of humanities research that actually is measurable, is not usually significant.
>
> Candidate 14 extract 4

There is no discussion of this difference within the script, such as an attempt to analyse which is more likely. It seems highly possible that this stems from the time-limited format of the CAEALT exam: that little time is available for true engagement with the ideas presented or to ensure that an argument is internally consistent with no clear holes. The time limitations, with limited time / options for redrafting, mean that a candidate must rely on essay and argumentation structures that they are already familiar with rather than to discover the most appropriate structure through writing and rewriting.

Critical engagement

CAEALT	Self assess.	
Argument: 6	3 (-0.71)	2.2.4 Identify the evidence which supports, confutes, or contradicts a thesis

The candidate does not demonstrate critical engagement with the sources beyond selection of relevant ideas: however, as noted earlier (section 5.2.3.3) there may be a tendency for the fact that sources are presented in the paper to lead the candidate to assume that the sources are innately sound; perhaps the CAEALT rubric needs to change to instruct the candidates that critical engagement is required. Alternatively, an element of source selection could be introduced to enforce evaluation. The difference between the CAEALT score and the candidate's self-assessment in this subskill may be because the CAEALT marking criteria do not place great significance on this area; subskill 4.2.1 (develop thesis convincingly) is more prominent at the upper end of the marking bands and given a self-assessment rating of 4 out of 5.

In conclusion, in this area the candidate is right to identify synthesis and critical engagement as weaker areas: She is able to summarise accurately, draw together ideas that support a coherent thesis and demonstrate good comprehension of complex texts but, at least in this snapshot, demonstrates limited critical engagement and little larger-scale synthesis of ideas or resolution of conflicts. However, as previously mentioned, the CAEALT is aimed at postgraduate entry. This is a lower level than the candidate is performing at, even in her weaker areas, and therefore there is no expectation that such weaknesses, unless extreme, will necessarily be reflected in her CAEALT scores.

5.2.4.2 Possible mismatches between CAEALT construct and AL taxonomy

CAEALT	Self assess.	
Engagement with sources: 6	3 (-0.71)	2.3.5 Understand separate ideas and then be able to see how these ideas form a whole
Engagement with sources: 6	2 (-1.86)	2.3.8 Synthesize information from several sources and incorporate it into a writing assignment

There are two issues of wording with subskills 2.3.5 and 2.3.8. First, subskill 2.3.5 does not specify whether it is about combining separate ideas within one source or across separate sources, while 2.3.8 does specify several sources. This suggests that, to remove this overlap, 2.3.5 should be reworded to target separate ideas in one source only (that this is a separate skill will be demonstrated below). Second, subskill 2.3.8 does not mention that ideas should be combined to form a coherent thesis, as is the case with 2.3.5, although this is the implication. Rewording would make this clearer.

There are also two wider issues surrounding synthesis under exam conditions. Firstly, it should be noted that synthesis traditionally takes place within pieces of writing written over a significantly longer period than this, as noted in the candidate responses to the test-taking experience questionnaire, with more time to select, understand and contrast sources. Additionally, it is usual for the title of the thesis to be decided upon by the writer, and thus to be the result of more personal engagement.

Secondly, I earlier discussed whether, for synthesis and critical engagement, it was necessary to present candidates with disagreements between sources to be resolved. A later iteration of the CAEALT may need to consider whether such contrast of source text is necessary for fully-realised synthesis across multiple sources; further discussion of this point takes place in section 6.1.3.

5.2.4.3 Candidate 5

Candidate 5 is 20 years old, female, and British. She does not speak other languages. She is in the 3rd year of an undergraduate degree in Education with English and Drama (although her end-of-second-year results cover English and Education only) at the University of Cambridge. The majority of her results are in the range 58-61 per cent; the exception was one of her Education modules where she scored 52, bringing her average score down to a high 2.2. Her previous highest qualifications were A-levels.

Candidate 5 is one of the three highest scoring undergraduate candidates, all of whom received an CAEALT score of 5.75; the only category for which she did not receive full marks was for coherence and cohesion, where she scored 5 out of 6. However, she scored the second lowest in university grades.

Coherence and cohesion

CAEALT	Self assess.	Tutor assess.	
Coherence & cohesion: 5	5	5	4.2.2 Organize information at both a section and paragraph level

Of the three paragraphs making up the body of the essay, none have an entirely clear focus. In each case, the topic sentence suggests a tighter focus than is given in the paragraph itself (paragraph 2 of the script presented in appendix 8.10.4 exemplifies this).

Topic sentence	Contents of paragraph in script
Before the works of scholars in the arts and humanities can be evaluated, it seems it must first be valued as a field of study. (paragraph 2)	Source texts [given in CAEALT] focus on the justification of the humanities The sciences don't have to justify themselves in the same way Humanities contribute to 'other fields' Humanities value is diverse and therefore less clearly evaluated Humanities scholars opposed to bibliometrics Value is less tangible for the humanities
Fisher (2000) notes that 'performance measures… narrow whereas the arts expand'. When humanities are evaluated using these narrow measures the subject can appear to lose some value. (paragraph 3)	Contrasting science and humanities may suggest that only one is useful Science is linear and humanities are 'expansive' Linear output may have more economic impact The subjects are very different, particularly for the humanities where subfields don't share criteria
It could then be argued that both the sciences and humanities should be evaluated using independent criteria. (paragraph 4)	Judgements of quality for humanities cannot be quantitative Usefulness change when viewed with different values Humanities is accessible to a wider audience

Table 20: candidate 5, topic sentences

Paragraph 2 is loosely grouped around the concept of value, but moves into areas of measurement (humanities is less clearly evaluated, bibliometrics). Paragraph 3 is about the narrowing effects of performance measurements, but, as with paragraph 2, moves towards areas of measurement. Paragraph 4 aims to lay out the claims of using independent criteria for the humanities and the sciences, then goes back to the content of paragraph 1 in discussing the value of the humanities.

In this subskill there is a clear difference between the performance in the CAEALT and the candidate's academic performance as reflected in the self- and tutor assessment. In fact, across subskills relating to all categories, the self- and tutor assessments rated coherence and cohesion the highest by a very small margin (0.2 marks).

In her feedback on similarities between the CAEALT and her everyday university life, candidate 5 wrote 'I would usually not write an essay so quickly after reading source material. I am used to preparing much more for a timed essay and considering sources' arguments etc.'. This suggests that the subskills that the candidate is missing are related to the ability to digest sources and assemble arguments at speed. These subskills are not included in RQ1's checklist, and it seems unlikely that they are contextually or cognitively relevant to the academic construct, seeming more akin to spoken debate or other specialised circumstances.

Engagement with sources

CAEALT	Self assess.	Tutor assess.	
Engagement with sources: 6	4	4	2.3.2 Make connections to related topics, information or prior knowledge, even when they are not obvious.

I will note briefly here that this subskill, which other candidates in the case studies have not demonstrated, is included by this candidate, indicating that it can be elicited by the CAEALT.

To give one example of the generation of ideas through synthesis, she notes that the thrust of the texts given are all essentially defending the humanities, that 'science subjects do not face the same criticisms, making any comparative methods instantly unequal' (paragraph 2). She continues to say that 'this positioning of the two subjects in conflict does perhaps is what inspires opinion that only one can be useful' (paragraph 3), a framing of the discussion that does not appear explicitly anywhere in the sources but is a legitimate contribution to the discussion.

As an English student, where engagement with texts is as texts rather than as sources of knowledge, such awareness of textual issues may be a fundamental skill in this subject and an assumed part of writing in this field.

5.2.4.4 Possible mismatches between CAEALT construct and AL taxonomy

CAEALT	Self assess.	Tutor assess.	

Engagement with sources: 6	3	5	4.1.1 Use of quotation, paraphrase and summaries to avoid plagiarism

When discussing candidate 7, I noted that, given the exam format, referencing requirements may not be apparent to the candidate. Candidate 5 did notice the referencing requirements as given in the rubric, and referenced well, but noted that ‘I would not normally pay so much attention to referencing in an exam context, as quotes, years etc would be memorised in advance’. This suggests both that the exam format does allow referencing to take place, but also that including referencing requirements in the rubric is not sufficient support in itself to ensure it. Further support seems necessary (possibilities for this are included in section 7.2.2).

6 Overall discussion

In our exploration of RQ2, the taxonomy produced in RQ1 has been shown to work, for the most part: there are a few subskills which require rewording to clarify the exact area targeted, and a few subskills which have the potential not to be elicited, even in high-scoring scripts, and a few which have not been observed as elicited in the CAEALT.

As a final stage, I will now review RQ1 in the light of insights gained in the process of investigating RQ2, by looking at the checklist of subskills to discuss issues of wording and revisit the analysis of the CAEALT to discuss which subskills were elicited in practice.

6.1 Validation of the RQ1 checklist

The quantitative and qualitative analyses carried out in RQ2 suggest that the RQ1 checklist as manifest in the CAEALT is a reasonable reflection of the skills needed for university success.

The quantitative analysis showed an overall correlation with university grades (τ=0.518, p=0.04, Cohen's d=8.42). No correlation was found for either of the other measures of university success: self- and tutor evaluation. I have suggested some reasons why this may be the case: firstly, there may be a difference in expectations as opposed to the CAEALT requirements: candidates' expectations of required performance may be higher or lower than actual required performance. Secondly, there is likely to be a link between how often a student receives effective feedback on a particular skill and the accuracy of their rating. It is also possible that there could be a difference stemming from exam technique or exam conditions: performance under exam conditions may not mirror real-life performance. Thirdly, self- or tutor ratings may have been consciously inflated through a desire to be seen as more competent (self-assessment) or to represent a student well / avoid being over-critical (tutor assessment). Finally, there is a difference in process vs product: difficulties a candidate has had in producing an essay may not be reflected in the final form of the essay as seen by the tutor. Thus, the fact that these measures of academic literacy do not correlate does not indicate that the RQ1 checklist is invalid.

6.1.1 Domain: Argument

In the quantitative analysis, argument correlated with university grades, although with a slightly higher p-value (τ=0.344, p=0.074, Cohen's d=8.4).

The qualitative analysis showed a range of performances from candidates, particularly in the areas of macro-structure of argument, where weaker candidates had a disconnect between the evidence, individual ideas and the overall thesis. In the case of candidate 7, there are smaller-scale structural issues: the flow of individual paragraphs can be unclear. This contrasts with candidate 14, whose arguments are complete and logically sound.

One key area of difference between the CAEALT task type and real-life university writing is that the candidate has been given an essay topic, rather than choosing one for themselves. This does not seem to have affected correlations under this category, but I will return to this under engagement with sources, below.

The case studies suggest that argument as tested in the CAEALT and as manifest in the RQ1 checklist may indeed correlate well with university mark in a larger-scale study, and further research in this area may well be worthwhile. However, my literature review noted that subject-specific practices do vary, something that has not been accounted for in this study due to small sample size. Any further research would need to take this area particularly into consideration.

6.1.2 Domain: Coherence and cohesion

In the quantitative analysis, a strong correlation was found between coherence and cohesion and university grade across the sample (τ=0.576, p=0.003, Cohen's d=8.43).

Coherence and cohesion was the marking category with the lowest average rating for the postgraduate candidates, but this is of limited relevance as all postgraduates scored either 5 or 6 in all categories, that is, the variety of marks was limited by being at the top of the marking scale.

Coherence and cohesion was not a key issue for any of the students in the case studies. One area that recurs across the case studies is a tendency to base the text structure on listing thematically-linked ideas in little depth rather than on exploring a few ideas in greater detail (subskill 4.2.1). This was a particular issue for candidate 10, but the scripts of the other case studies also followed this tendency to a lesser extent. This suggests that this subskill as defined in RQ1 is not well elicited in this particular synthesis-style task type i.e. one where there is little overlap of topic or opinion, or selection by the candidate of source material. It may be that if sources share the same topic and facts, but provide different interpretations of those facts, that this subskill will be demonstrated to better effect.

As with argument, candidates performed better at the paragraph level than the whole-text level. Signposting and paragraphing was performed reasonably well and, particularly in the case of candidate 10, the structure indicated by topic sentences was more coherent than the actual structure of the essay. These areas are often a key focus of academic English courses, which may imply that – in both argument and coherence and cohesion – whole-text structure is less taught before/in the early stages of an undergraduate course (which, if so, would have implications for the construct of an undergraduate entry exam), or that the student has less-concrete guidelines to establish whether they have been successful in this area.

6.1.3 Domain: Engagement with sources

The quantitative analysis found a correlation between engagement with sources and university grades (τ=0.467, p=0.014, Cohen's d=8.42).

The sample as a whole, as well as all four case studies, struggled with the category of engagement with sources. For the two weaker candidates, the key issue was excessive lifting and poor referencing. I discounted the idea that this may reflect a lack of knowledge of appropriate source use in the candidates, as it seems unlikely they could have completed one or two years of an undergraduate course without being aware of this. I then hypothesised that they considered such lifting acceptable given the exam conditions: that because the examiner is aware of the presented source material, that referencing may be unnecessary.

I also note here that the quantity of lifting demonstrated by the weaker case studies meant that they received a higher score for academic language use at the cost of engagement with sources. As the categories are equally weighted this is suitable in this case, as the score gain from a higher level of language will be compensated for by a lower mark for engagement with sources, but any exams with unequal weightings may wish to consider whether this is appropriate. It also seems likely that raters will need to be trained to be alert to lifting, and a clear policy given in rater-training documentation. I would also suggest the use of an electronic means of plagiarism detection to eliminate human error in this area.

Synthesis was another key issue: for both weak case studies the level of synthesis demonstrated was in stitching together arguments from the text and arranging them in themes in a more or less cohesive argument (a process having more in common with knowledge telling than knowledge transformation), rather than in critical analysis of sources to draw out relationships and highlight contradictions. At a higher level (candidate 14), issues were found with appropriate local and global synthesis. This candidate functioned at the same level as the weaker candidates in that her use of sources was primarily taking the ideas, choosing common themes and arranging the ideas (relatively untransformed) around these themes. Her higher mark reflected the fact that she was able to paraphrase and summarise highly successfully, as well as synthesise key ideas from individual sources on a more local scale.

The lack of synthesis in an otherwise highly-competent script suggests that true synthesis of sources may not be elicited under exam conditions (see below). This is a particular issue when taking into account Chan's 2013 cognitive construct of reading-into-writing, where connecting and generating and selecting relevant ideas are two different processes – the former process is not being elicited in the CAEALT. This advanced level of synthesis traditionally takes place within pieces of writing produced over a significantly longer period than this, with more time to select, understand and contrast sources. The timed format, and the reduction in processing time, means that non-obvious differences between sources may not be spotted by candidates; to alleviate this, it may be that contrasts between texts have to be apparent on a whole-text level (e.g. one text in favour, the other against).

It is also possible that this issue could be exacerbated because the writer is not choosing their own sources (and, as mentioned under Argument above, is not arriving at their own essay title through their review of these sources). A parallel could be drawn between this and a subskill that appeared on the original taxonomy but not on the final checklist: the subskill 'structure writing so that it moves beyond formulaic patterns that discourage critical examination of the topic and issues' was included in the taxonomy, scoring 12.5 (ICAS=5, NYSTE=2.5, Websites=5). Discovering a structure that serves the argument being conveyed rather than a formulaic template is a useful academic skill. However, I suggested earlier that the time-limited nature of the CAEALT means that the candidate may not have time to fully absorb the content of the sources and thus may not have properly integrated this content into the argument they present.

I also note here that the RQ1 checklist of subskills does not clearly differentiate between local and global synthesis. This will need to be remedied in a future iteration of such a checklist.

The lack of critical analysis, even in the stronger script, opens a discussion whether critical analysis can be expected under timed exam conditions, when sources are provided and no selection of source text is required. This should be viewed in the context of the comments from the test-taking experience questionnaire, where seven students noted that they would usually write essays over several days, which allowed greater thought and closer engagement with sources. Additionally, the fact that sources are provided may indicate to candidates that no critical engagement is required, as they are 'pre-approved'. These issues will require further research as to whether they are the case, and if so, how critical engagement can be operationalised. For the latter the solution may be as simple as explicit inclusion in the rubric. Alternatively, if candidates were instructed to only use a subset of the presented texts, this would allow the inclusion in the source material of obviously irrelevant texts.

To summarise the findings under the category of engagement with sources, this form of reading-into-writing task allows weaker and stronger candidates to engage with the text – albeit at the cost of unintentional plagiarism (and a decision must be made if the negative washback from this unintentional plagiarism would be sufficient to make this task type unattractive); this study presents no reason why this task type is not suitable for use at both undergraduate and postgraduate levels, as long as the weaknesses discussed above are considered.

6.1.4 Domain: Academic language use

Academic language use was the category where no correlation was found between university grades and CAEALT performance (τ=0.235, p=0.233, Cohen's d=8.39). I particularly note that academic language use was highly rated across all candidates and tutors, and well displayed in the CAEALT scripts. Additionally, all CAEALT candidates had previously passed a LP gatekeeping requirement before beginning their courses, whether that requirement was explicit (a language exam) or implicit (having previously studied in an English-speaking country). This uniformity of LP may provide a partial explanation to the quantitative finding that the academic language use marking category did not correlate with university grades, and it is possible that a sample taken from the larger applicant pool would show a correlation.

As previously mentioned, there is a tension between lifting from the source material and language accuracy, in that the more lifting there is from the source material, the less language accuracy can be demonstrated. For this reason, there is limited capacity to comment on the language of the two weaker case study candidates. Both do display rather mechanical (and in the case of candidate 7, occasionally misleading) use of linking phrases, with a tendency to start sentences with these, while the higher-level candidate demonstrated a greater range and flexibility of language.

In general, all four case studies, and the wider sample, all communicated successfully, with consistent use of a formal register, with very few examples of language-related incomprehensibility (candidate 10's 'scalable scholars' as one example). The errors that were found suggest that at this level, the ability to construct and explain on a larger scale than the clause is more of a concern.

6.1.5 Rewording of subskills

In the case studies, I suggested that three subskills required rewording to remove overlap and ambiguity:

Subskill number	Original wording	Rewording
1.1.2	Fully understand essay questions	Fully understand the task requirements of essay questions, including understanding of genre conventions, readership and wording of the rubric / task
2.3.5	Understand separate ideas and then be able to see how these ideas form a whole	Understand separate ideas within one source and see how these ideas form a whole
2.3.8	Synthesize information from several sources and incorporate it into a writing assignment	Understand separate ideas from several sources and see how these ideas form a whole

Table 21: subskills requiring rewording

Subskill 1.1.2 suffered from ambiguity in that it could be interpreted as either understanding the explicit, denotational meaning or as understanding the task requirements. As LP is not the focus of this checklist, the subskill now targets understanding of task requirements. The intention is that it targets understanding of genre expectations and awareness of the reader implied by the genre, but also appropriate argumentation structures for different essay question instructional words, such as 'identify', 'discuss', 'outline' and so on.

Subskills 2.3.5 and 2.3.8 have been distinguished from each other by making them target separate ideas in one source or across several sources respectively. In the case studies (candidates 7 and 14), text-level and intertextual synthesis of ideas was often demonstrated separately; they are also, according to Khalifa and Weir, separate cognitive processing levels (Khalifa and Weir, 2009, p.43) so this seems likely to be a more cognitively valid classification.

The reworded, final, checklist is presented in Table 12. Within a study-skills, EGAP, reading-into-writing academic literacy context, this seems likely to be an adequate representation of the academic literacy construct.

6.2 To what extent is the Cambridge Assessment English Academic Literacy Test representative of this construct?

In the context of the case studies, only some of the subskills in the construct were analysed in detail. Of those, some were not consistently elicited across the candidate scripts (Table 22). From this list of subskills, I hypothesised that the subskills in Table 23 are likely to suffer from the same issue, although they have not been analysed to the same depth.

1.1.2	Fully understand essay questions
5.2.1	Proofread to eliminate errors in grammar, mechanics and spelling, using standard English conventions
2.1.5	Identify authorial attitude
2.2.1	Identify the main thesis of a whole text
2.2.2	Determine major and subordinate ideas in a particular passage

2.3.5	Understand separate ideas within one source and see how these ideas form a whole

Table 22: subskills not consistently elicited in case studies

1.1.3	Duly consider audience and purpose
2.1.4	Decipher the meaning of vocabulary from the context
2.2.4	Identify the evidence which supports, confutes, or contradicts a thesis
2.2.5	Critically assess the authority and value of research materials that have been located
2.3.4	Understand 'rules' of various genres
6.2.1	Use revision techniques to improve focus, support and organisation

Table 23: subskills which seem likely not to be consistently elicited under test conditions

A few different categories can be seen here: those which are not elicited in a study-skills timed-conditions test model (write for different audiences and for different purposes, write appropriately in different genres, use revision techniques); those which may be taking place but are not observable (e.g. spot errors when proofreading own work, guess meaning of vocabulary) and those which do not appear to need to be elicited to produce an adequate or even a good essay (e.g. identify the attitude of the author, identify evidence which supports an author's thesis). These latter two may be surprising: I suggest that one of the potential drawbacks of the reading-into-writing task type alone is that reading comprehension cannot be closely targeted. That these subskills are not elicited is a consequence of this particular task type (reading-into-writing, under timed conditions, one task type only) rather than of the CAEALT's particular interpretation of this task type.

These categories raise the issue of the tension between the skills that are understood to be needed at university and those skills that can be demonstrated under exam conditions. This was briefly raised in the discussion of the NYSTE (section 4.1.3.2), which combined reading-into-writing with a multiple-choice reading comprehension component. Here it was noted that adding the MCQ reading comprehension component allows a fuller coverage of the academic construct, as subskills such as 'identify the attitude of the author', which may be otherwise hard to elicit (see above), can be targeted directly. I noted earlier that this introduced an element of contextual invalidity, which may be appropriate for the purposes of the NYSTE, but may not carry across to other academic tests. Whether or not the contextual invalidity of this approach counteracts the benefit of greater subskill coverage is an area requiring further investigation. The practical implications should also be noted: that another exam paper needs to be produced, and that extra time would be needed to take an extra multiple-choice component.

Further analysis should be considered before this checklist can be considered fully valid for test analysis. However, it is likely to serve well as an indication of an ALT's coverage.

Of the 25 subskills on the RQ1 checklist, 15 are likely to be elicited by the CAEALT (see Table 24). Comments on how the CE CAEALT could be amended to cover more of the checklist are in the conclusion.

Subskill number	Academic literacy subskills	
1.1.1	Generate ideas for writing by using texts in addition to past experience or observations	Y
1.1.2	Fully understand the task requirements of essay questions, including understanding of genre conventions, readership and instruction words	Y
1.1.3	Duly consider audience and purpose	NR
1.2.1	Structure writing so that it is clearly organized, logically developed and coherent	Y
2.1.2	Reading strategies: skim, scan, read for detail	Y
2.1.4	Decipher the meaning of vocabulary from the context	NO
2.1.5	Identify authorial attitude	NO
2.2.1	Identify the main thesis of a whole text	NO
2.2.2	Determine major and subordinate ideas in a particular passage	NO
2.2.4	Identify the evidence which supports or contradicts an author's thesis	NO
2.2.5	Critically assess the authority and value of research materials that have been located	NE
2.3.2	Make connections to related topics, information or prior knowledge, even when they are not obvious.	Y
2.3.4	Understand 'rules' of various genres	NR
2.3.5	Understand separate ideas within one source and see how these ideas form a whole	Y
2.3.8	Understand separate ideas from several sources and see how these ideas form a whol e	Y
3.1.1	Vary sentence structures and word choice as appropriate for audience and purpose	Y
3.1.2	Use vocabulary appropriate to college-level work and the discipline	Y
4.1.1	Use of quotation, paraphrase and summaries to avoid plagiarism	Y

4.2.1	Develop main point or thesis	Y
4.2.2	Organize information at both a section and paragraph level	Y
5.1.1	Link ideas to each other appropriately	Y
5.2.1	Proofread to eliminate errors in grammar, mechanics and spelling, using standard English conventions	NO
6.1.1	Develop thesis convincingly with well-chosen examples, reasons and logic	Y
6.2.1	Use revision techniques to improve focus, support and organization	NO
7.1	Provide essays	Y

Table 24: Analysis of CAEALT

Y = elicited

NR = Not relevant to a study-skills timed-conditions test

NO = May be taking place but are not always observable

NE = not necessarily elicited in a good answer

7 Conclusion

This study has worked towards the production of a taxonomy of subskills needed for academic reading and writing at university by drawing on the literature, university websites and existing ALTs. The checklist and the CAEALT were then compared to establish the extent to which the subskills are covered in this test. I have also made an initial exploration into the concurrent validity of the CAEALT, with the indication being that it correlates well with university grades.

The key findings of this thesis have been, first, that the three sources consulted are for the most part in agreement on the subskills they consider necessary for university performance, although there were disagreements in the priorities given to these subskills. Second, the results presented here suggest that the CAEALT may correlate with academic success to some extent. While the correlation was strongest for overall mark and for the marking categories coherence and cohesion and engagement with sources, it is possible that argument and academic language use will also correlate in a larger-scale study conducted with participants who have not been through university selection procedures.

The checklist of academic skills has been explored through case studies and I have suggested that some subskills are more likely to distinguish between candidates than others. Further research is needed to establish the validity of all subskills, and to see if they can indeed distinguish between different levels.

7.1 Limitations

The domain of this thesis is undergraduate academic university-level study, primarily within the context of university admissions. It has also been restricted to the social sciences and humanities as the literature indicates faculty expectations may be substantially different in other areas.

The main limitation of this study is the small sample size, both in absolute number and in variety of demographic. The numbers in the quantitative analysis are not sufficient to draw robust conclusions, and can be indicative only. The limited size of the sample also necessitated covering a wider range of subjects, institutions and academic levels than was originally intended, as well as a mix of modular and final-exam based courses. There was also a restricted range of LP (C1/C2 in cases where LP results were available) and of university grade (2.1 and 1sts, with two exceptions only). Opportunity sampling meant that a wider range of abilities, as manifest in university grades, could not be obtained. These limitations mean that conclusions drawn from the quantitative data must be indicative only. The qualitative case study analysis covers only some of the subskills, but has still allowed a more in-depth and meaningful perspective on the elicitation of the subskills presented in the checklist.

This study was initially intended to provide insight into the undergraduate admissions context, and so I began by recruiting only undergraduate students. After five months of recruiting it became clear that I would not be able to reach the required sample size by targeting undergraduate students alone, and so broadened the pool to include postgraduate students. This inclusion was a methodological compromise on two fronts: first, including two educational levels introduces new confounding variables (see section 2.2.2). Second, the taxonomy drawn up in RQ1 was primarily intended for undergraduate use, particularly as the websites consulted covered undergraduate skills only. Given the eventual inclusion of postgraduate students in this study, their performance would ideally have been compared against a taxonomy deliberately including the content of postgraduate study skills websites. However, it is interesting that the postgraduate case study (candidate 14, section 5.2.4.1) still showed that some aspects of her performance were lacking, suggesting that at least some of the subskills on the checklist apply to both undergraduate and postgraduate study.

Additionally, this study has assumed that grouping L1 and L2 together is not especially problematic (see section 3.2.2 for further discussion). Although the literature suggests that such grouping is acceptable, this is not an uncontroversial decision; further study may be necessary to establish that this is in fact a suitable assumption in the admissions test context.

The use of self- and tutor assessment, while a useful part of the methodology in creating other measures of success and thus potentially reducing confounding variables, is open to limitations in that such scores cannot be independently verified. While I have attempted to use these measures to triangulate onto the academic literacy subskills elicited by the CAEALT, this has been with limited success as no significant correlations appeared in the quantitative analysis, and conclusions from the case studies are exploratory in nature. A future study may wish to include a brief standardisation stage, such as aligning the Likert scale to university grades, prior to the students and tutors completing such assessments.

Finally, I acknowledge the limitations standard for a criterion-related validity study into university skills: that university coursework assesses much more than just academic literacy or LP. The methodology used has aimed to counteract this (through collection of alternative, more targeted, measures of success) and through the using lower τ values, but it is still likely to affect the results presented here.

7.2 Implications

7.2.1 Further research

A study of a similar design, but with both a larger sample size and a more controlled range of participants in terms of university subject and level of study, as well as a wider range of university marks and LP, is highly recommended. The LP variable is particularly important: I have hypothesised that a reason no significant correlation was found for academic language use was that the sample had already passed a gatekeeping LP test.

Three wider issues have been discussed by this thesis: first, that the field of academic literacy testing has not yet reached a consensus on the efficacy of testing subskills through tasks specifically targeted at these subskills (for example, the NYSTE's receptive, multiple-choice testing of particular subskills), versus the contextually-valid method of a reading-into-writing task alone (such as the TEEP). A future study comparing these two methods of subskill elicitation is recommended to see which is the most effective in predicting university success.

Following this, there is also a tension between the subskills that are required for academic writing at university and the extent to which those subskills can be tested under exam conditions. I have suggested that this is an issue for source use in particular, especially in the areas of synthesis of sources, lifting from given sources and critical engagement. Other areas that may prove problematic are proofreading, revision, spotting errors and subskills involving comprehension – all of which are necessary for university, but difficult to elicit in a reading-into-writing test. Investigation is necessary as to whether these areas do in fact need to be included in an academic literacy test; if so, whether alterations can be made to task specifications that will elicit them; if not, which alternative means of elicitation are possible.

There is also a lack of clarity in the literature on which subskills should be acquired pre-university, which are undergraduate and which postgraduate skills. The same is true of whether some subskills are taught within faculties or as more general study skills, and so whether they can fairly be included in a study-skills test. A future study may wish to compare this thesis' taxonomy with one or more EAP curricula, as this may clarify some of these ambiguities, as well as providing another clear list of subskills at a granular level.

Finally, the checklist proposed here is in need of full validation.

7.2.2 Suggestions for revision of the CAEALT

Overall, the CAEALT compares well with currently existing ALTs in terms of construct and subskill coverage.

Until further research exists into the most appropriate way of eliciting appropriate source use, this thesis suggests that the necessity for appropriate in-text citation of given source material, and of critical engagement, be clearly stated in the task rubric (Khalifa and Weir, 2009) but also exemplified clearly as this study suggests that simple inclusion in the rubric is not sufficient.

Eliciting synthesis of sources across texts is more difficult as it is a harder concept for candidates to grasp: even those who possess the targeted skill may not have encountered it as an explicit criterion or under exam conditions. I have also discussed the possibility that time limitations may prevent this subskill from being elicited at all, or may make the test less construct-specific (the ability to absorb and then construct arguments at speed forming no part of the proposed taxonomy in RQ1). If it remains necessary for these subskills to be tested under exam conditions, I suggest choosing source texts to include clear and easily noticeable disagreement between sources: this approach is likely to elicit synthesis, simply as contrasting views will need to be reconciled. This should also elicit some critical engagement; another area that has not been clearly elicited in the CAEALT.

An alternative to exam-assessment may be presented by a coursework or portfolio style assessment: by removing time limitation it seems likely that better inter-text synthesis and clearer arguments will be the result. This will also address some other subskills that are not covered by the CAEALT such as redrafting and rewriting.

8 Appendices

8.1 Taxonomy

Subskills are listed from highest to lowest scoring.

Academic literacy sub-skills	Competency number	Skill appears in final checklist	Marking category	ICAS	NYSTCE: ALST	TEEP	Test total	Dartmouth College	Open University	University of Kent	Website total	Total
Develop main point or thesis		Y	C	5	5	5	15	4	4	3	11	26
Proofread to eliminate errors in grammar, mechanics and spelling, using standard English conventions		Y	AL	5	5	5	15	3	4	4	11	26
Structure writing so that it is clearly organized, logically developed and coherent		Y	C	5	5	5	15	4	4	2	10	25
Organize information at both a section and paragraph level		Y	C	5	5	5	15	4	4	2	10	25
Read texts of complexity without instruction and guidance		COMBINED	AL	5	5	5	15	3	3	3	9	24
Develop thesis convincingly with well-		Y	A	5	5	5	15	4	4	1	9	24

chosen examples, reasons and logic												
identify the main idea of a text		Y	E	5	5	5	15	3	3	2	8	23
Fully understand essay questions		Y	A	0	5	5	10	4	4	4	12	22
Duly consider audience, purpose		Y	AL	5	5	5	15	2	4	1	7	22
determine major and subordinate ideas in passages		Y	E	5	5	5	15	3	3	1	7	22
make connections to related topics or information (or prior knowledge) even when they are not obvious.		Y	E	5	5	5	15	2	2	3	7	22
Synthesize information in discussion and written arguments		COMBINED	E	5	5	5	15	3	2	2	7	22
Vary sentence structures and word choice as appropriate for audience and purpose		Y	AL	5	5	5	15	2	4	1	7	22
Critically assess the authority and value of research materials that have been located		Y	A	5	5	2.5	12.5	3	4	2	9	21.5
Generate ideas for writing by using texts in addition to past experience or observations		Y	E	5	5	5	15	3	2	1	6	21
Understand separate ideas and then be able to see		Y	A	5	5	5	15	2	3	1	6	21

how these ideas form a whole												
Synthesize information from assigned reading		COMBINED	E	5	5	5	15	1	3	2	6	21
Synthesize information from reading and incorporate it into a writing assignment		COMBINED	E	5	5	5	15	2	2	2	6	21
Link ideas appropriately		Y	C	5	5	5	15	0	4	2	6	21
summarise information		COMBINED	E	5	5	2.5	12.5	3	2	3	8	20.5
Reading strategies: skim, scan, read for detail		Y	E	0	5	5	10	3	4	3	10	20
Use revision techniques to improve focus, support and organization		Y	C	5	2.5	2.5	10	4	4	2	10	20
Synthesize ideas from several sources		Y	E	5	5	5	15	0	3	2	5	20
Provide essays		Y	AL	5	5	5	15	2	3	0	5	20
Summarize ideas and/or information contained in a text		COMBINED	E	5	5	2.5	12.5	3	2	2	7	19.5
identify the evidence which supports, confutes, or contradicts a thesis		Y	A	5	5	5	15	0	3	1	4	19
decipher the meaning of vocabulary from the context		Y	AL	5	5	5	15	2	0	1	3	18
Use of quotation, paraphrase and summaries to avoid plagiarism		Y	E	0	5	5	10	0	4	3	7	17
Identify authorial attitude		Y	E	5	5	2.5	12.5	1	2	1	4	16.5
understand 'rules' of various genres		Y	AL	5	2.5	5	12.5	3	1	0	4	16.5

Draw conclusions from given reading		N	A	0	5	5	10	0	2	1	3	13
Structure writing so that it moves beyond formulaic patterns that discourage critical examination of the topic and issues		N	C	5	2.5	0	7.5	3	2	0	5	12.5
Report facts or narrate events		N	AL	5	0	0	5	2	4	0	6	11
Understand inference		N	E	0	5	5	10	0	0	1	1	11
Use vocabulary precisely to produce the given effect		N	AL	0	5	5	10	0	1	0	1	11
Identifying suitable excerpts of text for direct / indirect quotation		N	E	0	5	2.5	7.5	0	3	0	3	10.5
Understand and integrate quantitative data		N	E	0	5	0	5	2	3	0	5	10
Text types: research proposals, dissertations, literature reviews		N	AL	0	0	0	0	2	4	2	8	8
Discipline-specific writing		N	AL	0	0	0	0	2	4	0	6	6
selecting appropriate texts		N	E	0	0	0	0	0	3	3	6	6
Comparing and contrasting two (or more texts)		N	E	0	5	0	5	0	0	0	0	5
Anticipate possible counter-claims		N	A	0	5	0	5	0	0	0	0	5

8.2 First analysis of CAEALT

Y = Required for task	**5**
P = Possibly required for task	**2.5**
N = not required	**0**

Cognitive process	**Competency number**	**Academic literacy sub-skills**	**In CE ALT**	
Conceptualisation: task representation and macro-planning	1.1.1	Generate ideas for writing by using texts in addition to past experience or observations	P	2.5
	1.1.2	Fully understand essay questions	Y	5
	1.1.3	Duly consider audience and purpose	Y	5
Conceptualisation: revising macro plan	1.2.1	Structure writing so that it is clearly organized, logically developed and coherent	Y	5
Meaning and discourse construction: careful global reading	2.1.2	Reading strategies: skim, scan, read for detail	Y	5
	2.1.4	decipher the meaning of vocabulary from the context	Y	5
	2.1.5	Identify authorial attitude	Y	5
Meaning and discourse construction: selecting relevant ideas	2.2.1	identify the main thesis of a whole text	Y	5
	2.2.2	determine major and subordinate ideas in a particular passage	Y	5
	2.2.4	identify the evidence which supports, confutes, or contradicts a thesis	Y	5
	2.2.5	Critically assess the authority and value of research materials that have been located	Y	5
Meaning and discourse construction: connecting and generating	2.3.2	make connections to related topics, information or prior knowledge	Y	5
	2.3.4	understand 'rules' of various genres	P	2.5
	2.3.5	understand separate ideas and then be able to see how these ideas form a whole	Y	5
	2.3.8	Synthesize information from several sources and incorporate it into a writing assignment	Y	5

Translation	3.1.1	Vary sentence structures and word choice as appropriate for audience and purpose	N	0
	3.1.2	Use vocabulary appropriate to college-level work and the discipline	Y	5
Organising ideas in relation to input texts	4.1.1	Use of quotation, paraphrase and summaries to avoid plagiarism	Y	5
Organising ideas in relation to writer's own texts	4.2.1	Develop main point or thesis	Y	5
	4.2.2	Organize information at both a section and paragraph level	Y	5
Low-level monitoring and revising: editing while writing	5.1.1	Link ideas appropriately	Y	5
Low-level monitoring and revising: editing after writing	5.2.1	Proofread to eliminate errors in grammar, mechanics and spelling, using standard English conventions	P	2.5
High-level monitoring and revising: editing while writing	6.1.1	Develop thesis convincingly with well-chosen examples, reasons and logic	Y	5
High-level monitoring and revising: editing after writing	6.2.1	Use revision techniques to improve focus, support and organization	P	2.5
(Task types)	7.1	Provide short answer responses or essays	Y	5

8.3 Ethics committee approval form

UNIVERSITY OF BEDFORDSHIRE
Research Ethics Scrutiny (Postgraduate Research Students)
When completing this form please ensure that you read and comply with the following:
Researchers must demonstrate clear understanding of an engagement with the following:

1. *Integrity* - The research has been carried out in a rigorous and professional manner and due credit has been attributed to all parties involved.
2. *Plagiarism* - Proper acknowledgement has been given to the authorship of data and ideas. 3. *Conflicts of Interest* - All financial and professional conflicts of interest have been properly identified and declared.

4. *Data Handling* - The research draws upon effective record keeping, proper storage of date in line with confidentiality, statute and University policy.
5. *Ethical Procedures* - Proper consideration has been given to all ethical issues and appropriate approval sought and received from all relevant stakeholders. In addition the research should conform to professional codes of conduct where appropriate.

6. *Supervision* - Effective management and supervision of staff and student for whom the researcher(s) is/are responsible
7. *Health and Safety*- Proper training on health and safety issues has been received and completed by all involved parties. Health and safety issues have been identified and appropriate assessment and action have been undertaken.

The **Research Institutes** are responsible for ensuring that all researchers abide by the above. It is anticipated that ethical approval will be granted by each Research Institute. Each Research Institute will give guidance and approval on ethical procedures and ensure they conform to the requirements of relevant professional bodies. As such Research Institutes are required to provide the University Research Ethics Committee with details of their procedures for ensuring adherence to relevant ethical requirements. This applies to any research whether it be, or not, likely to raise ethical issues. Research proposals involving vulnerable groups; sensitive topics; groups requiring gatekeeper permission; deception or without full informed consent; use of personal/confidential information; subjects in stress, anxiety, humiliation or intrusive interventions must be referred to the University Research Ethics Committee.

Research projects involving participants in the NHS will be submitted through the NHS National Research Ethics Service (NRES). The University Research Ethics Committee will normally accept the judgement of NRES (it will never approve a proposal that has been rejected by NRES), however NRES approval will need to be verified before research can commence and the nature of the research will need to be verified.

Where work is conducted in collaboration with other institutions ethical approval by the University and the collaborating partner(s) will be required.

The **University Research Ethics Committee** is a sub-committee of the Academic Board and is chaired by a member of the Vice Chancellor's Executive Group, appointed by the Vice- Chancellor and includes members external to the University

Research Misconduct: Allegations of Research Misconduct against staff or post graduate (non-taught) research students should be made to the Director of Research Development.

October 2014

UNIVERSITY OF BEDFORDSHIRE

Research Ethics Scrutiny (Annex to RS1 form)

SECTION A To be completed by the candidate

Registration No: 1618186

Candidate: Martine Holland

Degree of: MA by Research

Research Institute: CRELLA

Research Topic: Cognitive and predictive validity of the Cambridge English Academic Literacy T est

External Funding: Course funded by Cambridge English Language Assessment

The candidate is required to summarise in the box below the ethical issues involved in the research proposal and how they will be addressed. In any proposal involving human participants the following should be provided:

- clear explanation of how informed consent will be obtained,
- how will confidentiality and anonymity be observed,
- how will the nature of the research, its purpose and the means of dissemination of

  the outcomes be communicated to participants,
- how personal data will be stored and secured
- if participants are being placed under any form of stress (physical or mental) identify

  what steps are being taken to minimise risk

  If protocols are being used that have already received University Research Ethics Committee (UREC) ethical approval then please specify. Roles of any collaborating institutions should be clearly identified. Reference should be made to the appropriate professional body code of practice.

October 2014

The proposed research will require participants to complete the Cambridge English Academic Literacy Test under exam conditions.

After potential participants have expressed an interest, they will be informed by email of the purpose of the research and of the measurements they will be asked to provide, in line with BAAL protocol. On the day of the research they will be presented with a hard copy of this information, including details of how this data is to be confidentially stored, and will be asked to sign a declaration authorising the use of their data.

Data collected is to be:

1. The results of the test
2. Self-reported end-of-year scores for the previous academic year (year 1)
3. Self-reported scores at the end of the coming academic year (year 2)
4. Self-assessment of performance in the test
5. Tutor feedback on the participant's academic literacy, in a checkbox

   format (collected at the beginning and end of year 2)

Data will be de-identified in the final thesis by the random assignment of candidate numbers. The removal of secondary identifiers / coding into broader categories may also be made necessary by the demographic of participants. If qualitative data is quoted, it will be attributed to an alias. Data will be stored securely and be destroyed five years after the conclusion of the research. It will be made clear to both the participant and their tutor that neither will be informed of the results of the data collected e.g. the tutor will not be told the participant's test results.

Participants may, in the time between taking the exam and being asked for end of year 2 scores, choose not to participate. It is expected that the year 2 information will cover a smaller number of participants to take this into account.

While taking an exam can cause some anxiety, the low-stakes nature of the exam and the completely confidentiality of the results will be emphasised.

Participants will initially be contacted via a regular bulletin that is sent out to all students in a certain department at the University of Cambridge, summarising the research and asking participants to contact us if interested in participating. This will require the initial participation of the relevant university administrators.

The extent of the financial inducement offered to candidates is yet to be decided, but is expected to be a voucher for no more than £10-15 in exchange for two and a half hours of their time.

October 2014

Answer the following question by deleting as appropriate:

1. Does the study involve vulnerable participants or those unable to give informed consent (e.g. children, people with learning disabilities, your own students)?

   No
   If **YES**: Have/will Researchers be DBS checked?

2. Will the study require permission of a gatekeeper for access to participants (e.g. schools, self-help groups, residential homes)?

   Yes

3. Will it be necessary for participants to be involved without consent (e.g. covert observation in non-public places)?

   No

4. Will the study involve sensitive topics (e.g. sexual activity, substance abuse)?

   No

5. Will blood or tissue samples be taken from participants?

   No

6. Will the research involve intrusive interventions (e.g. drugs, hypnosis, physical exercise)? **No**
7. Will financial or other inducements be offered to participants (except reasonable expenses)? **Yes**
8. Will the research investigate any aspect of illegal activity?

   No

9. Will participants be stressed beyond what is normal for them?

   No

10. Will the study involve participants from the NHS (e.g. patients) or participants who fall under the requirements of the Mental Capacity Act 2005?

    No

If you have answered yes to any of the above questions or if you consider that there are other significant ethical issues then details should be included in your summary above. If you have answered yes to Question 1 then a clear justification for the importance of the research must be provided.

*Please note if the answer to Question 10 is yes then the proposal should be submitted through **NHS research ethics approval procedures** to the appropriate **NRES**. The UREC should be informed of the outcome.

Checklist of documents which should be included:

(Tick as appropriate)

Project proposal (with details of methodology) & source of funding	Y
Documentation seeking informed consent (if appropriate)	Y
Information sheet for participants (if appropriate)	Combined with above
Questionnaire (if appropriate)	Y

October 2014

Applicant declaration

I understand that I cannot collect any data until the application referred to in this form has been approved by all relevant parties. I agree to carry out the research in the manner specified and comply with the statement of ethical requirements on page 1 of this form. If I make any changes to the approved method I will seek further ethical approval for any changes.

Signature of Applicant: ……….. ……………… Date: ……19/08/2017……………..

Signature of Director of Studies:

Date: ……08/08/17…………………….

This form together with a copy of the research proposal should be submitted to the Research Institute Director for consideration by the Research Institute Ethics Committee/Panel

Note you cannot commence collection of research data until this form has been approved

SECTION B To be completed by the Research Institute Ethics Committee:

Comments: Application approved via Chair's action.

Approved

Signature Chair of Research Institute Ethics Committee:

Date: 29/08/2017 *This form should then be filed on the student's record*

If in the judgement of the committee there are significant ethical issues for which there is not agreed practice then further ethical consideration is required before approval can be given and the proposal with the committees comments should be forwarded to the secretary of the UREC for consideration.

There are significant ethical issues which require further guidance

Signature Chair of Research Institute Ethics Committee: Date:

This form together with the recommendation and a copy of the research proposal should then be submitted to the University Research Ethics Committee

8.4 Research instruments

8.4.1 Student declaration of research permission

Student Declaration

The purpose of this test, which has already been extensively trialled, is to establish the extent to which performance on this test can predict future performance on undergraduate courses at the University of Cambridge, with the intention of using such a test for future admissions purposes or to gain diagnostic information for the information of tutors after admissions.

By taking part in this trial you are agreeing that the data listed below [including contact details and other personal data] you provide at this trial may be stored by the researcher in the UK and used for test development, research, and validation purposes.

As part of your participation in the research today the researcher will request: demographic data, information about your results in any previous language tests, your score in this test, a self-assessment of your skills in particular areas of academic performance, your self-reported scores for the academic year you have just completed, and information on your experience taking today's test.

At the end of the current academic year you will be asked to provide your self-reported scores for this year.

Your tutor will be approached at the same two points of the year to provide feedback on your performance. By taking part in this trial you also agree that your tutor may supply the researcher with this data unless you specifically request to opt out by notifying your Data Protection Officer in writing; you also agree that any data supplied by your tutor about you that falls under this agreement may be stored by the researcher in the UK.

The results of this trial will not be used in any way to assess your performance on any course or for entry to any course, and individual results will not be passed to your institution or shared with any third party outside Cambridge Assessment. Any published or publicly disseminated reports about this trial will be general in nature and individual students will not be identified in any reports arising from this trial.

This agreement shall be governed by and construed in accordance with the laws of England and Wales and the parties hereby submit to the exclusive jurisdiction of the English courts.

I have read and understood this agreement. I agree to participate in this trial and for all data, as specified above, to be used for test development, research, and validation purposes.

Signed: ____

Print Name: ____

Department: ____

Date: ______________________________

8.4.2 Demographic information

Candidate number ____________________________________

1. Gender ____________________________________
2. Age ____________________________________
3. Nationality ____________________________________
4. First language ____________________________________
5. Other languages spoken ____________________________________
6. Course enrolled on ____________________________________
7. College enrolled at ____________________________________
8. Highest previous qualification ____________________________________

9. Previous experience with language proficiency tests e.g. IELTS, TOEFL **Y / N**

   If yes, please give the date of the exam, the overall mark received and reading and writing test scores:

   ____

10. Please nominate a tutor to be contacted for data collection:

    ____

8.4.3 Questionnaire after taking test

Candidate number ____________

In what ways is the test you have just taken:

a) Similar to the reading and writing activities in your everyday university life?

b) Different from the reading and writing activities in your everyday university life?

8.4.4 Self-reported end of year scores

Candidate number:

Subject:

Department:

Please report your scores for the academic year you have just finished. If any scaling factors are present in the computation of your mark, such as exam results being scaled in line with coursework performance, please report this below.

Any scaling factors present:

Exam	Paper	Mark	Total possible score

The information you have reported will remain confidential. Individual information will not be shared with any third party outside Cambridge Assessment. Any published or publicly disseminated reports about this trial will be general in nature and individual students will not be identified in any reports arising from this trial.

8.4.5 Student self-assessment form

		1 Strongly disagree	2 Disagree	3 Neutral	4 Agree	5 Strongly agree	6 Don’t know	7 N/A
1	I find it easy to think of ideas to write about, using my own experience and any source texts provided							
2	I fully understand essay questions							
3	I am able to write for different audiences and for different purposes							
4	I am easily able to structure my texts in a coherent and well-developed way, connecting related ideas or information							
5	I am able to read texts quickly to get a general understanding of the text and to find relevant information							
6	I can easily guess the meaning of unfamiliar vocabulary from its context							
7	I can always identify the attitude/opinion of the author							
8	I can always identify the main thesis of a whole text							
9	I can always identify the major and subordinate ideas in a particular passage of text							
10	I find it easy to identify evidence which supports or contradicts a writer’s thesis							
11	When choosing source texts for an essay, I find it easy to assess the authority and value of research materials							
12	I am able to make connections between related topics, information or prior knowledge, even when they are not obvious							
13	I find it easy to form a coherent thesis from separate ideas							
14	I am comfortable writing appropriately in different genres							

15	I find it easy to combine ideas from several sources in one writing assignment							
16	I find it easy to choose the appropriate sentence structures for a particular purpose or audience							
17	I find it easy to choose the appropriate words for a particular purpose or audience							
18	I find it easy to use university level vocabulary							
19	I find it easy to appropriately quote, paraphrase and summarise another writer's views in my own work.							
20	I am able to express and develop the main point of my text							
21	I find it easy to organise information at a section or paragraph level							
22	I find it easy to link ideas to each other appropriately							
23	I find it easy to spot errors when I proofread my own work							
24	I am easily able to support my thesis convincingly with supporting evidence and logic							
25	I find it easy to revise my texts when I write							
26	I am confident when writing in the essay genre							

Any additional comments:

8.4.6 Tutor assessment form

		1 Strongly disagree	2 Disagree	3 Neutral	4 Agree	5 Strongly agree	6 Don't know	7 N/A
	Ideas							
1	The student can generate sufficient ideas in their writing, using both their own experience and any source texts provided							
2	The student draws appropriate connections between related ideas, information or prior knowledge, even when they are not obvious							
	Source materials							
3	The student consistently chooses source texts of appropriate authority and value							
4	The student can always identify the attitude/opinion of the author							
5	The student can always identify the main thesis of a whole text							
6	The student can always identify the major and subordinate ideas in a particular passage of text							
7	The student finds it easy to identify evidence which supports or contradicts a writer's thesis							
8	The student can easily integrate ideas from several source texts appropriately into their essay							
9	The student can consistently and suitably integrate quotation/summary/paraphrase into their writing							
	Thesis							
10	The student is easily able to form a coherent thesis from separate ideas							

11	The student consistently and convincingly supports his/her thesis with reasons, logic and well-chosen examples							

	Language and structure							
12	The student fully understands essay questions							
13	The student is able to write appropriately in different genres							
14	The student is fully confident writing in the essay genre							
15	The student can easily choose the appropriate sentence structures for a particular purpose or audience							
16	The student can easily choose the appropriate words for a particular purpose or audience							
17	The student can fully express and develop the main point of his/her text							
18	The student can organise information appropriately at a section or paragraph level							
19	The student can easily and appropriately link ideas to each other							
20	The student is easily able to structure his/her texts in a coherent and well-developed way							
21	The student gives no appearance of struggling with comprehension of less common vocabulary							
22	The student can easily use the vocabulary required at university level							
23	The student does not submit work with proofreading errors							

Any additional comments:

8.5 CAEALT, self-assessment and tutor assessment scores

Candidate number		ALT overall	ALT Argument	ALT C&C	ALT Academic language use	ALT Engagement with sources	University grades	Self-assess. overall	Self-assess. Argument	Self-assess. C&C	Self-assess Academic language use	Self-assess. Engagement with sources	Tutor assess. overall	Tutor assess. Argument	Tutor assess. C&C	Tutor assess. Academic language use	Tutor assess. Engagement with sources
1		4.5	6	5	4	3	63.80	3.42	4.00	2.20	3.50	3.78	3.50	4.00	2.75	4.00	4.00
2		5.75	6	6	5	6	67.25	3.88	3.75	3.80	3.75	4.11	4.26	4.75	4.25	4.11	4.13
3		4.5	4	5	4	5	71.20	3.46	3.50	3.60	3.75	3.11					
5		5.75	6	5	6	6	58.60	4.19	3.50	4.40	4.38	4.22	4.83	4.75	5.00	4.78	4.88
6		5.75	6	6	6	5	67.60	4.23	4.25	4.20	4.38	4.11	4.78	4.75	4.75	4.56	4.88
7		3.5	3	4	4	3	61.00	4.46	4.00	4.60	4.50	4.56	4.61	5.00	4.75	4.89	4.13
8		4.75	6	3	6	4	62.50	4.00	4.00	4.20	3.88	4.00					
9		5.5	5	5	6	6	66.25	4.12	3.75	4.20	4.13	4.22					
10		3	3	3	4	2	40.63	4.31	4.50	4.00	4.13	4.56					
11		4	4	3	6	3	65.75	4.12	3.75	4.00	4.13	4.33					
12		6	6	6	6	6	81.00	4.69	4.25	5.00	4.75	4.67					
14		6	6	6	6	6	89.00	3.63	3.50	3.60	3.86	3.50					
15		6	6	6	6	6	67.00	4.08	3.50	3.80	4.38	4.22					
16		5.5	5	5	6	6	68.75	4.15	4.00	4.20	4.25	4.11					

17		5.75	6	6	5	6	70.50	4.04	4.00	4.00	4.00	4.11					
19		5.5	6	5	6	5	72.50	4.92	5.00	5.00	5.00	4.78					
20		6	6	6	6	6	85.00	4.50	4.75	4.20	4.50	4.56					
22		5.25	5	5	6	5	66.89	3.62	3.50	3.20	3.38	4.11					

8.6 Candidate university grades

Candidate number	Mark	Weighting	Total possible score	Average mark
1	62	Equal	100	63.8
	65	Equal	100	
	68	Equal	100	
	56	Equal	100	
	68	Equal	100	
2	67	Equal	100	
	70	Equal	100	
	66	Equal	100	
	66	Equal	100	67.25
3	70	Equal	100	71.2
	70	Equal	100	
	72	Equal	100	
	75	Equal	100	
	69	Equal	100	
5	61	Equal	100	58.6
	58	Equal	100	
	61	Equal	100	
	52	Equal	100	
	61	Equal	100	
6	63.3	Equal	100	67.6333333
	72.6	Equal	100	
	67	Equal	100	
7	62	Equal	100	61
	60	Equal	100	
	61	Equal	100	
	61	Equal	100	
8	60	Equal	100	
	65	Equal	100	62.5
9	64	Equal	100	66.25
	66	Equal	100	
	68	Equal	100	
	68	Equal	100	
	62	Equal	100	
	66	Equal	100	
	68	Equal	100	

	68	Equal	100	
10	40	Equal	100	40.625
	40	Equal	100	
	45	Equal	100	
	40	Equal	100	
	40	Equal	100	
	40	Equal	100	
	40	Equal	100	
	40	Equal	100	
11	71	Equal	100	65.75
	64	Equal	100	
	68	Equal	100	
	60	Equal	100	
12	77	Equal	100	
	84	Equal	100	
	78	Equal	100	
	91	Equal	100	
	86	Equal	100	
	77	Equal	100	
	77	Equal	100	
	90	Equal	100	
	81	Equal	100	
	77	Equal	100	
	73	Equal	100	81
14	9	Equal	10	
	10	Equal	10	
	9	Equal	10	
	7	Equal	10	
	9	Equal	10	
	9	Equal	10	
	9	Equal	10	8.9/10
15	65	Equal	100	
	65	Equal	100	
	65	Equal	100	
	75	Equal	100	
	65	Equal	100	67
16	74	Equal	100	
	71	Equal	100	
	68	Equal	100	
	62	Equal	100	68.75
17	79	15	100	
	73	30	100	

	69	15	100	
	74	30	100	
	66	15	100	
	56	15	100	70.5
18	75	Equal	100	75
19	68	Equal	100	
	82	Equal	100	
	72	Equal	100	
	70	Equal	100	
	73	Equal	100	
	75	Equal	100	
	65	Equal	100	
	80	Equal	100	
	70	Equal	100	
	72	Equal	100	72.7
20	85	Equal	100	85
22	66.89	Equal	100	66.89

8.7 Self-assessment raw scores

Candidate	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q26
1	4	4	4	2	2	2	4	3	4	3	4	5	4	4	4	4	5	3	4	2	2	3	5	5	2	1
2	5	4	3	4	4	4	4	4	4	3	4	4	3	3	5	3	4	4	4	4	4	3	5	4	4	4
3	4	4	4	4	2	4	2	2	4	4	2	4	2	4	4	4	4	4	4	4	4	4	2	4	2	4
4	5	4	5	4	5	5	4	4	4	3	4	4	3	4	5	5	5	5	5	5	5	5	5	4	4	5
5	5	4	5	4	5	4	4	5	4	4	3	4	3	5	5	5	5	5	3	5	5	4	2	3	4	4
6	5	4	4	4	3	5	4	3	4	4	5	4	4	4	5	4	5	5	5	4	5	4	4	4	4	4
7	4	4	5	4	5	3	5	5	4	5	3	4	4	5	5	5	5	4	5	4	5	5	4	4	5	5
8	4	5	4	5	4	4	4	4	4	4	3	4	3	4	4	4	3	4	5	4	4	4	3	4	4	5
9	4	4	5	5	5	5	4	3	4	4	4	5	4	4	4	3	3	5	5	3	4	5	4	3	4	4
10	5	4	3	4	5	5	5	5	4	4	5	4	4	4	5	4	4	4	4	3	4	4	5	5	5	4
11	4	4	5	4	5	4	4	4	4	4	4	5	4	5	5	4	4	4	4	4	4	4	3	3	4	4
12	5	5	5	5	5	5	4	4	5	4	4	5	4	5	5	4	5	5	5	5	5	5	4	4	5	5
14	3	4	5	4	5	5	4	4	4	3	3	dk	3	2	2	3	3	4	3	4	4	3	5	4	3	dk
15	4	3	5	3	5	5	4	4	4	3	4	5	4	5	5	5	5	3	3	4	4	5	3	4	3	4
16	5	5	4	3	3	5	4	4	4	3	4	4	4	4	4	4	3	4	5	4	5	5	5	4	4	5
17	4	4	4	4	5	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
18	5	5	4	4	5	5	4	4	5	4	4	5	4	4	5	4	4	4	4	5	5	5	3	4	4	4
19	5	5	5	5	5	5	4	5	5	5	5	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5
20	4	5	4	4	5	4	4	4	4	4	5	5	5	5	5	4	4	5	5	4	4	4	5	5	5	5
22	4	4	4	3	5	5	4	4	4	3	4	4	3	3	5	4	4	3	4	3	4	4	1	3	2	3

8.8 Tutor assessment raw scores

Tutor assessment	**1**	**2**	**5**	**6**	**7**
Q1	4	4	5	5	5
Q2	4	4	4	5	4
Q3	4	5	5	5	6
Q4	4	4	5	5	4
Q5	4	4	5	5	3
Q6	4	4	5	4	4
Q7	4	4	5	4	5
Q8	4	4	5	5	5
Q9	4	4	5	5	4
Q10	4	5	5	5	4
Q11	4	5	4	5	4
Q12	4	5	5	5	5
Q13	N/A	4	5	6	6
Q14	3	4	4	5	4
Q15	3	4	5	4	5
Q16	3	4	5	4	5
Q17	4	5	5	4	5
Q18	2	4	5	5	5
Q19	3	4	5	5	5
Q20	2	4	5	5	4
Q21	dk	4	4	5	4
Q22	dk	4	5	5	5
Q23	2	5	5	4	5

8.9 Comments from test-taking experience questionnaire

8.9.1 Similarities

Category	**How was the CAEALT similar to reading and writing activities in your everyday university life?**
Argumentation	the arguments [illegible] in the [illegible] were presented in a way similar to the [illegible] I do
	Evaluate question
	Argument focussed question that involved making a thesis and supporting arguments
	Having to write in an argumentative style
	The test is similar in terms of the kind of texts I had to deal with (which are generally similar to politics and sociology articles in style). It also resembles the type of quite analytical questions that HSPS students have to answer every week (the 'to want extent' in particular).
	Similar to exam format in terms of the structure of the essay
	Question very broad to encourage creative thinking
	Provided with a question/statement and asked to evaluate so you are able to write an argument but it doesn't lead you down a certain route
	Able to use your own argument and thought - no right or wrong
Critical engagement	Very similar - I am required to consult a wide range of resources, evaluate their usefulness and use them to construct a strong and well evidenced argument
	Evaluation of multiple resources
	Evaluation of the validity and strength of paper
	Need to evaluate and cross-compare several texts
	Use multiple sources to evaluate/cite/compare/argue against
Integrated	Very similar - I am required to consult a wide range of resources, evaluate their usefulness and use them to construct a strong and well evidenced argument
	Using my own ideas / interpretations alongside the work of scholars
	The writing part was quite similar: having time, seeing the readings, taking notes.
	Having to express my own opinion on a certain topic with using evidence from other resources
	Given an unfamiliar topic and asked to conduct research on it in the hope of writing a completed essay
	You need to use the texts to support your argument and conclusions

	It was similar in that it required a synthesis of a range of material, and also covered concepts that were new and unfamiliar. The texts were relatively dense and needed careful reading
	Reading different sources on a related topic and synthesising the relevant information into an essay along with your own opinion = very similar
Output	Essay length estimate is the same as university exam
	Writing formal pieces of essay writing
	Very similar. Most of the modules on my Criminology course have had an essay approach to assessment. I have only completed two exams throughout the three years
	Styple of writing
	Genre of writing
	Having a word count is similar, sticking to the word count is an important part of the tasks we're given at uni.
Selecting	Selecting the most relevant bits of information and identifying what is not important
	Reading large sums of information and having to pick out snippets that are important to the essay question
	The essay is similar to what I do in my everyday university life in that I have to select information and manipulate it in order to make a case for something. The range of reading text types is also similar to what I do.
Sources	The reading is similar in the sense that the style of writing of the authors is the same
	Some of the source materials provided were similar to what I may read when researching an essay. The process felt like a mix of a timed exam and a supervision essay - perhaps because of the unknown element
	They are similar in character, but not in length and time and complexity
	The citations are needed
	Evaluation of multiple resources
	Reading large sums of information and having to pick out snippets that are important to the essay question
	The test is similar in terms of the kind of texts I had to deal with (which are generally similar to politics and sociology articles in style). It also resembles the type of quite analytical questions that HSPS students have to answer every week (the 'to want extent' in particular).
	The essay is similar to what I do in my everyday university life in that I have to select information and manipulate it in order to make a case for something. The range of reading text types is also similar to what I do.
	Now I am writing my PhD it is very similar, although I found it difficult in not being able to quote externally.

	Use multiple sources to evaluate/cite/compare/argue against
	Academic literature
	It was similar in that it required a synthesis of a range of material, and also covered concepts that were new and unfamiliar. The texts were relatively dense and needed careful reading
	Reading different sources on a related topic and synthesising the relevant information into an essay along with your own opinion = very similar
Timed	The timed element and structure of the question were similar to previous exams
	Past exams I have done included sections for writing an essay in a limited time. So I was quite used to this. I thought it was good that three texts were different in layout, especially text C which had a table. Although it wasn't completely clear what the right-hand side table was trying to convey.
Misc	Given an unfamiliar topic and asked to conduct research on it in the hope of writing a completed essay
	The Question style is similar to questions given in exams and essay [illegible] in the sense of how they are phrased
	Not at all similar

8.9.2 Differences

	How was the CAEALT different from reading and writing activities in your everyday university life?
Integrated	Usually exams are closed-book (no texts allowed)
	I can't remember ever having to do a test which involved reading three large texts before writing my answer. It was a bit challenging at first because you didn't necessarily feel like you had time to really come to grips with the material. I ended u using a lot of my own pre-existing knowledge or experience to make up my answer. Rather than reading in detail, I tended to skim for things that would support my arguments.
Knowledge	I would usually have some pre-existing knowledge of the topic and the important scholars in the field
	Exam essays are different due to the memorizing of names and dates involved
	I write essays on topics that are more relevant to my subject
	Not a subject/question/theme familiar with
Output	Supervision essays are much longer and more complex (including the readings) and the whole process of reading is different.
	800 - 2000 words
	Shorter than the 2-3000 word papers I'm writing

	Word count of 800 is quite succinct - challenge to do!
Sources	The approach to using sources is different to how they are used in philosophy as we [illegible] assess the arguments.
	Different to the philosophical approach as it seems to demand more from the sources
	Supervision essays are much longer and more complex (including the readings) and the whole process of reading is different.
	I rather read experimental papers
	Usually I rely less on critical material and more on my own thoughts, but this is because I will have researched more / had more time to research the topic of the essay.
	Usually more substantial articles (and more of them) would be used
	In historical elements of Politics 1 main 'primary text' is used and then 'secondary texts' are used in relation to it
	Also the level of analysis of the first text and the language used is far more abstract than what I read in my university life.
	Less onerous citation requirements
	no bibliography - texts are provided
	The three formats of the given texts were different to what I am used to, especially the third one.
	Not being able to use a dictionary/internet was authentic to the uni exam context, but inauthentic to regular coursework/essay writing. Same goes for doing it by hand.
Sources provided	Sources provided to reference and inform argument which is not usually the case in university exams
	Generally there's more choice over which texts to choose to talk about when writing an essay (students select ones they think look interesting from a list of lots!)
	Selected sources (only 3) and quite similar - nothing too contradictory/controversial
	Being provided with sources is less similar - often going out and finding the right papers to draw on is part of the task. However, in the non-honours part of undergrad we were sometimes referred more directly to sources for answering an essay question.
	No reading on the topic on my own, to find an angle I'd like to explore more. I also have never written an essay with the source material chosen for me. In exam settings, it was based on what we had studied, and for essay assignments, we did our own research.
Time	The time used for reading is a lot longer to [illegible] comprehend arguments in university as I'm not trained to quickly read and understand material, but more to [illegible]
	Time provided was much more than university exams
	I usually write essays over a 3 - 7 day period, rather than in timed conditions (however my end of year exams are similar to this)

	I would usually not write an essay so quickly after reading source material. I am used to preparing much more for a timed essay and considering sources' arguments etc. I would not normally pay so much attention to referencing in an exam context, as quotes, years etc would be memorised in advance (and I would not use page numbers!)
	I take more time for writing the essay and finding resources (but on exams it takes roughly the same time)
	Longer time frame - have up to a week usually to write an essay, so more time for planning, writing, proof-reading etc.
	Time pressure
	Essays aren't usually timed (unless exam)
	The main difference is time pressure and the inability to consult other resources for clarification/amplification of the ideas. My MA course was assessed on coursework only, with no timed **** [illegible]. Also the lack of assessment criteria with this exam makes it a bit harder to know what's expected - I wasn't sure, for example, how much of my own voice or opinions I should bring in beyond the three source texts. Finally it is a very long time since I've had to handwrite so much text!
Topic	The topic is very different
	Topic wise, naturally
Typed	Handwritten vs typed
	The test was different because of having to hand write - all of my essays are done on the computer and so this was quite different! Other than that - I find it harder to plan handwritten essays, Text C wasn't something I would often encounter. Furthermore, in our essays defining key terms is key, so I guess this is somewhat different from this essay.
	The main difference is time pressure and the inability to consult other resources for clarification/amplification of the ideas. My MA course was assessed on coursework only, with no timed **** [illegible]. Also the lack of assessment criteria with this exam makes it a bit harder to know what's expected - I wasn't sure, for example, how much of my own voice or opinions I should bring in beyond the three source texts. Finally it is a very long time since I've had to handwrite so much text!
	Not being able to use a dictionary/internet was authentic to the uni exam context, but inauthentic to regular coursework/essay writing. Same goes for doing it by hand.
Misc	I plan my essays more carefully
	The main difference is time pressure and the inability to consult other resources for clarification/amplification of the ideas. My MA course was assessed on coursework only, with no timed **** [illegible]. Also the lack of assessment criteria with this exam makes it a bit harder to know what's expected - I wasn't sure, for example, how much of my own voice or opinions I should bring in beyond the three source texts. Finally it is a very long time since I've had to handwrite so much text!
	Often we'd be given a choice of 3-5 questions to answer so only having one option is a bit different. That's not always the case though.

8.10 Case study scripts

Asterisks indicate illegible text.

8.10.1 Candidate 7 script

Traditionally, humanities and sciences have been comparison points to each other, with occasionally inviting opposition (Small, 2013). There is a lively ongoing debate among scholars about using the same criterion system in assessing the validity and impact of research papers from the two disciplines. The bibliometric measures currently employed are used in practice for evaluating the validity and social impact of papers, and evaluating this into, for example, bases of university funding systems in several countries (Hug, Ochsner, and David, 2001). The issue with this application has been thought to be the wide array of differences in the types of social impact elicited, and the research methods used as sources for the different papers. Several authors have suggested that a new form of bibliometric measures should be used for humanities in order to provide a valid picture of the papers' quality, since the current system has significant overlaps with the one used in natural sciences. In this essay I will introduce the pros and contras of using the same criteria in evaluating humanities and science papers, and conclude that based on the available literature, the most beneficial solution would be to create separate guidelines for the two disciplines.

The pro side of the argument claims that it is a wrong approach to distinguish humanities from sciences, as it would reduce the value of the former one. Small (2013) extensively wrote about his arguments on how humanities have specific value and purpose for the society as a whole. He argues that research in the field makes a significant contribution to the economy and indirectly benefit the growth of GDP. This is measurable in terms of the income produced by bookshops, museums, heritage sites, theatres etc. Therefore, applying the same evaluative criteria as before is useful so that the economic impact can be distinguished between science and humanities research. Small's next argument is quite weak, although it can still rationalise why the criteria should remain. He claims that humanities facilitate the undertaking of happiness and hence research can be employed in the education system to raise more content adults. How this can be aided by the research evaluation criteria staying the same that humanities papers that stand out in this way of employability will be more visible for the public.

Altogether, all over the paper he presents quite holistic arguments for the value of humanities, from contributing to happiness to the discipline being 'needed by democracy'. Small admits that some of the claims are weak, but this doesn't diminish the value of humanities, and his point that their social impact is still very important, and it could provide a ground for why the current criterion system of papers should stay the same both in the humanities and sciences.

The arguments of the contra side of evaluating the two disciplines based on the same guidelines are noticeable stronger than that of the pro side. Scholars on this opinion suggest that a new system of evaluation should be created for humanities research. One of the strongest arguments comes from the fact that the evaluation system derives from classifying natural sciences papers (Vec, 2008). Therefore, grading research that wasn't done based on equations and formulas will not give a reliable measure of quality and quantity. (Academics Australia, 2008). In addition, multiple sources have also claimed vastly different approaches to research in the two disciplines, with different philosophies used during conduction of studies as well (Lack, 2008; Olmos-Penuela, Beneworth and Castro-Martinez, 2015). On the other hand, some humanities scholars do not deny quantifiability, but they still deem this practice unnecessary, as these indicators communicate information that is already widely known.

The second argument against using the same criteria is that since citation counts are also included there is a tendency to favour spectacular research and neglect ones from more marginalised fields. Another problem supporting this argument is the fact that authors often use self-citation or cite friends exclusively and this manipulate reliability (Charle, 2009).

The third line of argument claims that even if the overall evaluation is consistent throughout disciplines, within and between them the standard for what is acclaimed as a valid and important paper might differ noticeable (Herbert and Vaube, 2008).

Finally, Oimiss-Penuela, Benneworth and Castro-Martinez (2015) systematized the differences between humanities and sciences, and their finding also rather suggest a revival of validity guidelines. They proposed that humanities research doesn't need as much external validity as sciences do, since applicability of the results is relatively smaller. In addition, they cite Cassity and Aug (2006) who wrote that humanities are less related to business innovation, and the authors also claimed that there is less demand for humanities research than for science research.

In conclusion, the debate and the controversy is still very much ongoing, but in my opinion most of the findings seem to suggest a small extent to which it is useful to evaluate humanities and science research. Of course, it is important to remind ourselves that the findings and the arguments do not decrease the value of humanities, and Small's (2013) arguments do support this on a certain level, showing that a holistic approach is important in understanding this and the societal impact cannot only be measured through how much *** a new *** innovation made. With a new system of research classification the practical application based on impact can be reliably done by recognising and utilising the difference between science and humanities.

8.10.2 Candidate 10 script

There are various reasons for why the study of arts and humanities is an impart value. Firstly, the study established by (Small, 2013) has found that the humanities have a 'contribution to make to our individual and collective happiness'. The aim of this essay is to establish whether it is possible to evaluate the work of scholars in the arts and humanities. Furthermore, the second aim is to discuss whether the work of those in the sciences can be evaluated using the same criteria.

Without a doubt, the studies of arts and humanities is questioned by other sciences. There are claims that humanities are less valuable to society than sciences. According to (Cassity and Ang, 2006) 'human research is less directly related to business innovation and is more a nice addition than critical success'. Furthermore, there is a stance that humanities scholars dedicate their time to the idea of 'blue-skies research' (Geelbrandsen and Kvik (2010) and (Hughes & Kitson, 2012).

Nevertheless, the stylised facts of the differences between the societal value of humanities and sciences are vague. While business researchers would claim that 'humanities researchers experience a lower demand for their research than is correspondingly the case for science researchers (Olmos-Penuela, 2015), scalable scholars would argue that the rate of involvement with national users in comparison to international users is greater for humanities researchers than for science researchers. (Olmos-Penula, 2015) It is difficult to evaluate the work of (Olmos-Penuela, 2015) due to the fact that the research of the author is factualized.

(Hug, Ochsne and Daniel, 2014) have conducted a study into the quality and criteria of research within the humanities. They have established that the methods of research originate from the natural sciences. Furthermore, there is a fear of the negative 'steering effects' of indicators and a lack of consensus on quality criteria. There are also strong ~~quantifications~~ reservations against quantification. Nevertheless, some researchers claim that 'bibliometric indications are not well-suited to determine the quantity and quality of humanities research' (Archambault et al, 2016) Some would argue that the consensus regarding the criteria for good and bad research is non-existent (Herbert and Kaube, 2008). This is problematic, because it is difficult to evaluate the work of scholars in the arts and humanities if the research has no approval and therefore is invalid.

The value of humanities has been examined by (Small, 2013). There are five claims established. The first is that the value of humanities is meaningful since they study the meaning-making practices of the culture. Secondly, there is a significant pressure on how governments commonly understand use and prioritize the scale of economic usefulness. (Small, 2013) Thirdly, (Small, 2013) takes stance that the humanities have a contribution to make to our general happiness. Furthermore, the fourth claim 'democracy needs us' is the most ambitions argument now regularly heard for the humanities in Britain. The final claim is that the humanities matter for their own sake. (Small, 2013) The five arguments have been influential in ancient history and maintain persuasive power. It is an easy task to evaluate the work of (Small, 2013), since the scholar's publication is of significantly large content, in comparison to (Olmos and Penuels, 2013).

In conclusion, it is possible to evaluate the work of scholars and humanities. Howeve r, there are various factors that can affect a researcher's ability to evaluate such research. The content, as observed ~~in~~ between publications of (Small, 2013) and (Olmos-Penuela et al, 2015) is of significant importance, as well as the lack of consensus on quality criteria established by (Hug, Oschner and Dniel, 2014). Overall, evaluating the work of scholars is a task that can be evaluated to a great extent.

8.10.3 Candidate 14 script

It is an undeniable fact that the work of scholars is of paramount importance in deepening our knowledge of the world and anything we interact with. Both the humanities and the sciences conduct research projects with the aim of improving our quality of life and our awareness of the world around us. The ways in which this research can be evaluated has always drawn attention and the reason for this may be that evidence obtained from this evaluation is usually used to base decisions on; decisions about a specific research project or research in a specific field is worth investing in, thus leading to more research opportunities – and funding – becoming available. But is it fair or valid to try and apply the same evaluation criteria to the word of scholars in the arts and humanities as to the work of scholars in the sciences, and to what extent? This essay will argue that it isn't either fair or valid to try to evaluate these two areas in the same way, because of the differences between them and the non-suitability of the current criteria proposed. Firstly, I will discuss what differentiates the humanities from the sciences, with regards to assessing them. Secondly, I will discuss the reasons why the methods proposed so far cannot be applied to evaluate research in the humanities and finally I will conclude with some considerations on what steps need to be taken for the efficient assessment of humanities research.

Small (2013) states that the humanities are marked by their 'distinctive character'. Even though it is common for the humanities to be compared to the sciences, it is also interesting to realise that they encompass the sciences in a way, as without the humanities, we wouldn't be able to perceive knowledge, let alone analyse it, understand it and build on it. Like Small (2013) one can argue that the humanities study 'the meaning-making practices of the culture' and this is one of the reasons why, while Humanities research relates to smaller scales when compared to sciences research (Olmos-Penuela et al, 2015), there is value in promoting the former and strive for its fair evaluation. This smaller scale to which humanities research usually relates also means that the profile of the Humanities research users is very different from the profile of science research users. Humanities researchers work more directly with a broad range of users, who come mainly from the public and voluntary sectors and, more often than not, this is limited to a national level, while science researchers work mainly with firms and more often on an international level (Olmos-Penuela et al, 2015). These facts may hide the explanation to why, according to Olmos-Penuela et al

(2015), demand for humanities is lower than that for sciences research and also why humanities researchers 'rank lower than science researchers in formal economic impact indicators'. This could also justify the widely spread belief that the humanities are not as valuable as the sciences. Advocates of the value of the humanities and their impact on societies have argued that the benefits from humanities research can be translated not, into something intangible and vague, but also into measurable goods, such as increase in GDP and increase of growth for 'the economy proper' (Small, 2013), through the promotion of cultural activities. Overall, the humanities deserve to be evaluated in equal measure to the sciences because they contribute to societies just as much, only in different ways.

The second reason why humanities research cannot be evaluated using the same criteria as the ones used to evaluate science research lies in the methods that have been put forward so far. Most of these methods have been borrowed from the natural sciences (Hug et al, 2014), which renders them unsuitable. This is due to the non-linear fashion in which humanities research progresses and also the more evident fact that a lot of humanities research cannot be easily quantified. What scholars stress is that the part of humanities research that actually is measurable, is not usually significant and that indicators typically used to quantify research impact provide little new information to the assessor. Furthermore, because of the easy-to-manipulate nature of some indicators, such as citation counts, there always lies a risk of evaluation results that are skewed and which do not reflect the significance of all research projects, especially in the case of lesser researched fields (Hug et al, 2011).

Last but not least, it is proving difficult to apply the same criteria as those applied to science research evaluation projects, simply because the humanities lack shared criteria on quality. This discrepancy is not only apparent in the comparison of the humanities to the sciences but also within sub-disciplines of the humanities. Hug, et al (2014) state that criteria used in humanities research are not formalised, probably as a result of the scope of humanities research being primarily 'local' rather than 'international' as pointed out by Olmos-Penuela et al (2015).

To sum up, I believe it is quite clear that the work of scholars in the arts and humanities cannot be evaluated using the same criteria as the ones used for the work of scholars in the sciences because of the significant differences in traits of the two areas of research work and additionally because of the lack of an appropriate set of criteria and evaluation method, which would be fit-for-purpose, able to capture the essence of humanities research in order to evaluate it in a fair, reliable and valid manner.

8.10.4 Candidate 5 script

Scholars of the arts and humanities are frequently compared, with their work often being measured to a similar set of criteria. I will argue that while it is possible to evaluate these works using the same criteria, the practise is not productive and is damaging to both subjects individually. It is worth noting that this argument predominately focuses on the issues surrounding humanities within this topic. This is firstly as I write from the perspective of an arts student but also in response to the literature surrounding the question, in which a defense of the humanities and their validity is a recurring theme.

Before the works of scholars in the arts and humanities can be evaluated, it seems it must first be valued as a field of study. Small discusses a 'justification for the humanities' (2013, p.3) from the outset and this automatic defense of the subject introduces a key issue within the question, that humanities scholars are in constant defence of their subject. A problematic factor of then comparing Arts and humanities with the sciences is that science subjects do not face the same criticisms, making any comparative methods instantly unequal. Small goes on to argue that humanities studies, while perhaps not competing with the 'economic usefulness' (2013, p.3) of other subjects, provide a significant contribution to other fields. They prese nt a 'pluralistic account of value' (2013, p.3) which suggests that the measure of humanities studies lies in multiple elements that are less clearly evaluated. While extensive, Small's evaluation is qualitative and descriptive, providing little if any academic response to back up the insightful points made, however their writing remains valuable. One example is Hug, Oshsner and Daniel (2014), who note how humanities scholars are opposed to bibliometric measuring of their work and highlight how measurable output is not important in the humanities' (p.7), emphasising the way that within humanities studies a value is placed on features that are less tangible and more developmental, however still holding this validity within the work.

Fisher (2000) notes that 'performance measures… narrow whereas the arts expand'. When humanities are evaluated using these narrow measures the subject can appear to lose some value. Small (2013) discusses how the sciences and the arts and humanities are historically compared and even sometimes opposed and this positioning of the two subjects in conflict does perhaps is what inspires opinion that only one can be useful. Across the literature (Small, 2013; Crossick, 2009; Malas-Gallart, 2015) there is a focus on 'economic usefulness' (Small, 2013), a restrictive measurement that asks two fields with very different focuses to compare. Lack (2008) highlights that humanities scholars have a different way of processing – 'an expansion of knowledge' (p.14) rather than the more linear and measurable output that may be found in science research and the economic impact that this may make. It is also worth noting the difficulties with setting evaluation criteria for subjects that are so different, particularly within humanities where, as Herbe rt and Faube note criteria isn't 'transferable to other sub disciplines' (2008, p.40, translated in Hug et al (2004)) within the field itself. If the subsections of the humanities cannot be evaluated alongside each other, it is highly unlikely that the whole field of research can be actively compared to the sciences; another considerable field of study.

It could then be argued that both the sciences and humanities should be evaluated using independent criteria. Academic Australia (2008) emphasise the idea that the issue with evaluating humanities subjects is not that judgements of quality 'cannot be made but more that humanities cannot be given 'the quantitative expression' that the sciences are so well suited to (2008, p.1). Therefore there are perhaps benefits to assessing, the work of scholars in both fields differently. Olmos-Penuela et al (2015) highlight the difference of interpretation well, showing how usefulness may change when values are shifted, for example from economic to welfare. Their data can be viewed to suggest that humanities research is 'less valuable' (p.10) than science, as not so economically viable (Cassity & Ang, 2006). However, it is contrastingly apparent that humanities research works with and is more accessible to a wider audience and range of users (Hughes et al, 2011 and Oimos-Penuela at al, 2013a).

In conclusion, humanities scholars and scholars of science can be evaluated to the same criteria, however this will most likely never fully appreciate the benefits of each subject. Ultimately I would argue that neither is less worthy, however, they both have the potential to appear best when evaluated using specific and appropriate criteria.

9 Bibliography

Atkinson, R.C. and Geiser, S. 2009. Reflections on a century of college admissions tests. *Educational Researcher*, 38(9). Pp. 665-676.

Blue, G.M. 1988. Individualising academic writing tuition. In: P.C. Robinson, ed. 1988. *Academic Writing: Process and Product. ELT Documents 129.* Reading: British Council. pp. 95-99.

Bridgeman, B., Cho, Y. and DiPietro, S. 2016. Predicting grades from an English language assessment: the importance of peeling the onion. *Language Testing*. 33(3), pp. 307-318.

Bridgeman, B., McCamley-Jenkins, L. and Ervin, N. 2000. Predictions of freshman grade-point average from the revised and recentered SAT1: Reasoning Test. [Online]. New York: College Entrance Examination Board. Available from http://files.eric.ed.gov/fulltext/ED446593.pdf [Last accessed 20th April 2017].

Chan, S. 2013. Establishing the validity of reading-into-writing test tasks for the UK academic context. Unpublished PhD thesis: University of Bedfordshire.

Chan, S., Wu, R. and Weir, C. 2014. Examining the context and cognitive validity of the GEPT Advanced Writing Task 1: A comparison with real-life academic writing tasks.

LTTC-GEPT Research Report. Available from www.lttc.ntu.edu.tw/lttc-gept-grants/RReport/RG03.pdf

Chernyshenko, O.S. and Ones, D.S. 1999. How selective are psychology graduate programs? The effect of the selection ratio on GRE score validity. *Educational and Psychological Measurement.* 59(6), pp. 951-961.

Cho, Y. and Bridgeman, B. 2012. Relationship of TOEFL iBT scores to academic performance: Some evidence from American universities. *Language Testing*. 29(3), pp. 421-442.

Cotton, F. and Conrow, F. 1998. An investigation of the predictive validity of IELTS amongst a group of international students studying at the University of Tasmania. *IELTS Research Reports volume 1.* Pp. 72 – 115.

Cumming, A., 2013. Assessing integrated writing tasks for academic purposes: promises and perils, *Language Assessment Quarterly*, 10(1), pp. 1-8.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U. and James, M. 2006. Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL. *TOEFL Monograph Series MS-30.* Available from https://www.ets.org/Media/Research/pdf/RR-05-13.pdf

Dartmouth College. 2017. *What academic paper?* [Website] Available at: http://writing-speech.dartmouth.edu/learning/materials/materials-first-year-writers/what-academic-paper [Last accessed 15 June 2017].

Open University. 2017. *Skills for OU study.* [Website] Available from: http://www2.open.ac.uk/students/skillsforstudy/ [Last accessed 19th August 2017].

Dartmouth College. 2017. *Academic Skills Centre*. [Website] Available from: https://students.dartmouth.edu/academic-skills/learning-resources/learning-strategies [Last accessed 15th August 2017].

Dooey, P. 1999. An investigation into the predictive validity of the IELTS test as an indicator of future academic success. In N. S. K. Martin & N. Davison, eds, *Teaching in the disciplines / Learning in Context: Proceedings of the 8th Annual Teaching Learning Forum*. Perth: UWA., pp. 114–118.

Feast, V. 2002. The impact of IELTS scores on performance at university. *International Education Journal* 3(4), pp. 70–85.

Fyfe, M., Devine, A. and Emery, J. 2017. The relationship between test scores and other measures of performance. In: Cheung, K., McElwee, S. and Emery, J. 2017. *Studies in Language Testing 49: Applying the socio-cognitive framework to the BioMedical Admissions Test.* Cambridge: CUP. pp 143-180.

Gardner, S. 2011. Perspectives on the disciplinary discourses of academic argument [online]. *Corpus Linguistics Conference. Birmingham: Birmingham University.* Available from http://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-71.pdf

Harsch, C., Ushioda, E., Ladroue, C. 2017. Investigating the Predictive Validity of TOEFL iBT® Test Scores and Their Use in Informing Policy in a United Kingdom University Setting. *ETS Research Report Series,* 2017(1). Available from https://onlinelibrary.wiley.com/doi/full/10.1002/ets2.12167

Humphreys, P., Haugh, M., Fenton-Smith, B., Lobo, A. and Walkinshaw, I. 2012. Tracking international students' English proficiency over the first semester of undergraduate study. *IELTS Research Reports Online*, Available from https://www.ielts.org/-/media/research-reports/ielts_online_rr_2012-1.ashx

ICAS. 2012. Academic Literacy: A statement of competencies expected of students entering California's public colleges and universities. Sacramento: ICAS.

Ingram, D. & Bayliss, A. 2007. IELTS as a predictor of academic language performance (Part I). *IELTS Research Reports Vol. 7.* pp. 1-68.

Jordan, R. (1997). English for Academic Purposes: A Guide and Resource Book for Teachers. Cambridge: CUP.

Kerstjens, M., and Nery, K. 2000. Predicative validity in the IELTS test: A study of the relationship between IELTS scores and students' subsequent academic performance . *IELTS Research Reports Vol 3.* Available from https://www.ielts.org/-/media/research-reports/ielts_rr_volume03_report4.ashx

Khalifa, H. and Weir, C.J. 2009. Studies in Language Testing 29: Examining Reading: research and practice in assessing second language reading. Cambridge: CUP.

Khalifa, H., and Barker, F. (eds), 2015. *Research Notes, issue 62.* Cambridge: Cambridge English Language Assessment.

Khalifa, H. and Weir, C.J. 2009. Studies in Language Testing 29: Examining Reading: research and practice in assessing second language reading. Cambridge: CUP.

Knoch, U., Sitajalabhorn, W. 2013. A closer look at integrated writing tasks: towards a more focussed definition for assessment purposes. *Assessing Writing* 2013(18) pp. 300-308.

Lea, M.R. and Street, B.V. 1998. Student writing in higher education: an academic literacies approach. *Studies in Higher Education* 23(2) pp. 157-172.

Light, R.L., Xu, M., & Mossop, J. 1987. English proficiency and academic performance of international students. *TESOL Quarterly*, 21(2), 251–261.

Lillis, T. 2003. Student writing as 'academic literacies': Drawing on Bakhtin to move from 'Critique' to 'Design'. *Language and Education* 17(3), pp. 192-207.

McManus, I.C., Dewberry, C. et al. 2013. Construct-level predictive validity of educational attainment and intellectual aptitude tests in medical student selection: meta-regression of six UK longitudinal studies. *BioMed Central,* 11(243). pp. 1-21.

Milanovic, M. and Weir, C.J. 2004. Studies in Language Testing 18: European language testing in a global context. Cambridge: CUP.

Murray, N., 2016. Standards of English in Higher Education: Issues, Challenges and Strategies. Cambridge: CUP.

Nesi, H. and Gardner, S. 2006. 'Variation in disciplinary culture: university tutors' views on assessed writing tasks' In: Kiely, R., Rea-Dickins, P. et al (eds) *Language, Culture and Identity in Applied Linguistics*. London: Equinox Publishing Ltd. pp. 99-117.

Nesi, H. and Gardner, S. 2012. Genres across the disciplines. Cambridge: CUP.

New York State Education Department. 2014. Academic Literacy Skills Test (ALST) Test design and framework. Available from https://www.ccny.cuny.edu/sites/default/files/ltrcenter/upload/ALST-TestDesign-Framework.pdf [Last accessed on 20th February 2017].

Open University. 2017. *Skills for OU study.* [Website] Available from: http://www2.open.ac.uk/students/skillsforstudy/ [Last accessed 19th August 2017].

Patterson, R., Weidman, A., 2013a. The refinement of a construct for tests of academic literacy. *Journal for Language Teaching*, 47(1): pp 124-151.

Patterson, R., Weidman, A., 2013b. The typicality of academic discourse and its relevance of constructs of academic literacy. *Journal for Language Teaching* 47(1): pp 107-123.

Plakans, L. 2010. Independent vs. integrated writing tasks: a comparison of task representation. *TESOL quarterly,* 44(1). pp. 185-194.

Pollitt, A. 1988. Predictive Validity. In: Hughes, A., Porter, D., and Weir, C. (eds) *ELTS Validation Project: Proceedings of a conference held to consider the ELTS Validation Project Report*. pp. 62-66.

Postman, R.D. (4th ed) 2015. *NYSTE.* Barron's Educational Series: New York.

Rex. L.A. and McEachen D. 1999. If anything is odd, inappropriate, confusing, or boring, it's probably important: the emergence of inclusive academic literacy through English classroom practices. *Research in the teaching of English*, 24, pp. 65–129.

Rosenfeld, M., Leung, S., and Oltman, P.K. 2001. The Reading, Writing, Speaking, and Listening Tasks Important for Academic Success at the Undergraduate and Graduate Levels. *TOEFL Monograph Series* (MS-21): Available from https://www.ets.org/Media/Research/pdf/RM-01-03.pdf

Sawaki, Y., Nissan, S. 2009. *Criterion-Related Validity of the TOEFL iBT Listening Section*. ETS. Available from https://www.ets.org/research/policy_research_reports/publications/report/2009/hvea

Sebolai, K. 2016. Distinguishing between English proficiency and academic literacy in English. *Language Matters*, 47(1), pp. 45-60.

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

University of Kent, 2017. *Study guides.* [Website] Available from https://www.kent.ac.uk/learning/resources/study-guides.html [Last accessed on 27 April 2017].

University of Reading. 2017. *Test of English for English for Educational Purposes (TEEP)*. [Website] Available from: https://www.reading.ac.uk/ISLI/TEEP--english-language-test/islc-teep-practice-tests.aspx [Last accessed 03 August 2017].

Ushioda, E., and Harsch, C. 2011. *Addressing the needs of international students with academic writing difficulties: Pilot Project 2010/11. Strand 2: Examining the predictive validity of IELTS scores.* Warwick: Centre for Applied Linguistics, University of Warwick. Available from https://warwick.ac.uk/fac/soc/al/research/groups/llta/research/past_projects/strand_2_project_report_public.pdf

Weigle, S. 2011. *Validation of Automated Scores of TOEFL iBT Tasks Against Nontest Indicators of Writing Ability*. ETS. Available from https://www.ets.org/research/policy_research_reports/publications/report/2011/isty

Weir, C.J. (unpublished) The construct of academic literacy: a preliminary discussion paper. Unpublished: Cambridge English.

Weir, C. J. 1983. Identifying the Language Problems of Overseas Students In Tertiary Education In The UK, Unpublished PhD thesis: University Of London.

Weir, C.J., Chan, S. and Nakatsuhara, F. 2013. Examining the criterion-related validity of the GEPT Advanced Reading and Writing tests: Comparing GEPT with IELTS and real-life academic performance. *LTTC-GEPT Research Reports RG-01.* Taipai: The Language Training and Testing Center.

Yen, D. and Kuzma, J. 2009. Higher IELTS score: higher academic performance? The validity of IELTS in predicting the academic performance of Chinese students, *Worcester Journal of Learning and Teaching,* (3), pp. 1-7.