EAP TEACHER ASSESSMENT LITERACY

Thesis submitted for the degree of Doctor of Education at the University of Leicester

by

Anthony Manning MA BA

School of Education

University of Leicester

2013

ACKNOWLEDGEMENTS

Many thanks to Glenn Fulcher for all his guidance and supervision during the course of my studies at Leicester.

Thank you also to all the staff at Leicester who have marked and advised on my work. I would also like to extend my appreciation to the administrative staff of the Doctoral Studies Office in the School of Education for their kind support and assistance over the years.

Above all, I would like to send my sincere gratitude to Amy and Imogen who have put up with my absences, while I have been absorbed in MBW.

ABSTRACT

TITLE: EAP Teacher Assessment Literacy
NAME: Anthony Manning

The first chapter of this thesis begins with an introduction to the background associated with the research project and an overview of significant issues in English for Academic Purposes (EAP) and its assessment. This leads on to an explanation of the need for EAP Assessment Literacy. It is clarified, through reference to the research questions, that this study has been undertaken in order to investigate the extent to which the views of EAP teachers, on the topic of EAP testing and assessment, reflect language testing research and practices which comprise Assessment Literacy. Statement of the research questions and the hypothesis under investigation also explains that the thesis aims to ascertain how EAP Assessment Literacy can be sustained or enhanced.

Chapter 2 of the study presents the Literature Review which examines the conceptual frameworks surrounding Assessment Literacy and the context of EAP. The views of key proponents of Assessment Literacy are discussed alongside a review of associated research in the fields of Education, Applied Linguistics and EAP. Other resources and mechanisms available for the enhancement of Assessment Literacy through formal training or in-service learning opportunities are also discussed.

In the third chapter, the Research Design and Methodology, the researcher's philosophical approach is rationalised and an appropriate set of research methods for this enquiry into EAP Assessment Literacy is identified. For this particular study the choice of a pragmatic paradigm is clarified. A mixed methods approach is also described in addition to the design of a questionnaire and an interview protocol for data collection purposes.

Chapter 4 presents the Findings section of the research. Features of the research participant population are described and, in the case of the questionnaire, a method for filtering the total number of respondents to provide a more representative group is outlined. The results from the two main research tools are then presented in terms of key sets of data. These data sets include different question types and investigate the key areas of Assessment Literacy in EAP.

Through Chapter 5, Analysis and Discussion, the meaning revealed within the Findings chapter is interpreted further. This will serve to identify key information relating to EAP teachers' views on EAP testing and assessment practices and the extent to which these reflect language testing research practices which comprise Assessment Literacy.

A synthesis of the main outcomes of the research is then provided in Chapter 6, Conclusions and Recommendations. Suggestions regarding what needs to be undertaken in order to sustain or enhance EAP teacher Assessment Literacy are also provided in the recommendations. In addition, the limitations of the study are also acknowledged, so that the scope of the research is appropriately contextualised.

In Chapter 7, a list of references is provided detailing resources which have been cited during the research. Finally, an Appendix is included in Chapter 8 which includes the full questionnaire and the full set of interview questions. Access is also provided to transcripts from each of the interviews, the series of open question responses from the questionnaire and number of charts and diagrams which are not included in the main body of the thesis.

KEY WORDS

Assessment Literacy

English for Academic Purposes

LIST OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

TABLE OF CHARTS

GLOSSARY & ABBREVIATIONS

AALL	Association for Academic Language Learning
ALTE	Association of Language Testers in Europe
BALEAP	The British Association of Lecturers in English for Academic Purposes
BOS	Bristol Online Surveys
BrunELT	Brunel English Language Test
BWPT	BALEAP Working Party on Testing
CALL	Computer Assisted Language Learning
CALS	Centre for Applied Language Studies
CBT	Computer Based Test/Testing
CEFR	Common European Framework of Reference
CELTA	Certificate in English Language Teaching to Adults
CPD	Continuing Professional Development
CPE	(Cambridge) Certificate of Proficiency in English
DALT	Durham Academic Language Test
DELTA	Diploma in English Language Teaching to Adults
DVASs	Discrete Visual Analogue Scales
EAP	English for Academic Purposes
EGAP	English for General Academic Purposes
ELBA	English Language Battery
ELT	English Language Teaching
ELTS	English Language Testing Service
EPTB	English Proficiency Test Battery
ESAP	English for Specific Academic Purposes
ESOL	English for Speakers of Other Languages
ESP	English for Specific Purposes
ETS	Educational Testing Service
HEIs	Higher Education Institutions
Ho	The Null hypothesis
iBT	Internet Based Test/Testing
IDP	International Development Program Education Australia
IELTS	International English Language Testing System
ILTA	The International Language Testing Association
ISLC	International Study and Language Centre
KITE	Kent International Test of English
MA	Master of Arts
MCQs	Multiple Choice Questions
MPhil	Master of Philosophy
OELPT	Online English Language Proficiency Test
PGCE	Post Graduate Certificate in Education
PgCert	Postgraduate Certificate
PhD	Doctor of Philosophy
PTE Academic	Pearson Test of English Academic
QAA	Quality Assurance Agency for Higher Education
QTS	Qualified Teacher Status
SELMOUS	Special English Language Materials for Overseas University Student Groups
SELT	Secure English Language Test
TDA	Teacher Development Agency for schools

TEAP	Teaching English for Academic Purposes
TEAP	Test of English for Academic Purposes
TEAPP	Teaching English for Academic and Professional Purposes
TEEP	Test of English for Educational Purposes
TEFL	Teaching English as a Foreign Language
TELAS	Test for the English Language Assessment Service
TESOL	Teachers of English to Speakers of Other Languages
TLU	Target Language Use
TOEFL	Test of English as a Foreign Language
TSE	Test of Spoken English
TWE	Test of Written English
UELT	University English Language Test
UKBA	United Kingdom Border Agency
USELT	University of Sussex English Language Test
USEPT	University of Sheffield English Proficiency Test
VOIP	Voice Over Internet Protocol
WELT	Warwick English Language Test
WYSIWYG	What You See Is What You Get

CHAPTER 1 INTRODUCTION AND BACKGROUND

1.1 Chapter introduction

The focus for this thesis has emerged through the researcher's professional interest in Assessment Literacy, particularly in the context of English for Academic Purposes (EAP). This interest was first initiated as an EAP teacher and has since grown due to a developing career which now involves the management of a University language centre and responsibility for teaching, learning and assessment. In brief, the research conducted in this project is driven by a keen objective to investigate and enhance the professional practice of EAP teachers involved in testing and assessment.

Throughout this study the term EAP will be used to refer to the strand of English for Specific Purposes (ESP) which deals with teaching English with the aim of assisting learners' study or research in that language (Hyland, 2006, p.1; Flowerdew & Peacock, 2001, p.8; Jordan, 1997, p.1). The terms 'testing' and 'assessment' will also be used interchangeably in this thesis as umbrella terms for both formal and informal procedures associated with gathering language data for the purpose of evaluation (Davies et al., 1999, p.11). Whilst assessment is often accepted as broader in nature and is considered to encompass a wider range of evaluative methods, formal and informal, summative and formative; some individuals consider that testing describes a narrower more summative and evaluative activity, relating to scheduled, time-limited assessments undertaken under restrictive conditions (Allison, 1999, p.5-6). For the purpose of this study, the terms will be used synonymously with testing not being confined to formal modes of assessment. The view will be maintained that irrespective of which term is used, the act of evaluating and

measuring language ability requires the same level of responsibility regarding purpose, methodology and justification (ibid, p.6).

The researcher's interest in EAP testing and assessment first arose through the process of professional practice as an educator and assessor in the field of EAP. Initially, the need to engage with research arose through a personal requirement to cultivate a better understanding of good practice in EAP assessment design and test score interpretation. This situation has since progressed to an interest in investigating EAP teacher Assessment Literacy in the wider EAP sector. It is hoped that this will assist in assuring the quality of assessment procedures, so that stakeholders, in particular test takers themselves, are not disadvantaged. With this aim in mind, the goal of this inquiry is to identify trends in areas of EAP teacher assessment practice which will enable the evaluation of EAP teacher Assessment Literacy, along with the identification of recommendations and mechanisms for its sustenance and/or enhancement. It is anticipated that this will generate a number of potential applications which can be disseminated to EAP practitioners and which will have a practical and positive impact on the wider EAP teaching and assessment sectors.

1.2 The research problem and research questions

In order to anchor the research which the thesis will focus on, it is useful, at this stage, to refer to the research questions and accompanying hypothesis which underpin this particular project.

The research questions which drive this study are described below in Table 1:

Table 1 Thesis research questions

Research Questions
• To what extent do EAP teacher views on EAP testing and assessment practices reflect language testing research and practices which comprise Assessment Literacy?
• How can EAP Assessment Literacy be sustained or enhanced?

The hypothesis which has been formulated for the purpose of this research is:

EAP teachers who are involved in EAP testing and assessment have certain identifiable development requirements with regard to their knowledge and ability to implement assessment good practice and recommendations stemming from research.

In order to explore the above theory, the null version (H_o) of this hypothesis will be tested to see if it can be rejected and therefore if support is available for the original hypothesis itself. The null hypothesis is as follows:

EAP teachers who are involved in EAP testing and assessment do not have any identifiable development requirements with regard to their knowledge and ability to implement assessment good practice and recommendations stemming from research.

1.3 Significant issues, challenges and ideas

EAP has progressed from a marginal position to become a fervent source of empirical research in the field of English language teaching as well as a major factor in international student success in tertiary education (Hyland & Hamp-Lyons, 2002, p.2-3). Confirmation of this situation has been recorded over the course of the last five decades through publications such as Cowie and Heaton (1977), Robinson (1980), Widdowson (1983), Swales (1985) Jordan (1997), Flowerdew and Peacock (2001), Hyland (2006) and Guse and Thornbury (2011) which provide an account of the many and varied approaches, research findings, controversies and increasing levels of expertise which now support EAP teaching contexts.

In recent years, the theme of assessment has been growing in importance in all areas of education, not least in language education. As the number of students studying in international contexts through a medium other than their native language has increased rapidly, so the assessment of language as a medium for study in the form of EAP has also become increasingly significant (Blue et al., 2000, p.7). Consequently, assessment can be considered as a prevalent practice associated with the teaching and learning of EAP.

In 2012 BALEAP, The British Association of Lecturers in English for Academic Purposes, celebrated the 40th anniversary of its foundation (BALEAP, 2012b). Since the organisation's initial launch, as SELMOUS in 1972 (Jordan, 1997), significant innovations have been realised in the form of materials design and the establishment of fora for the sharing of good practice (Blue et al., 2000, p. 43). However, despite the growing global importance of EAP, until relatively recently there has been very little practical guidance for test and assessment development associated with English for Specific Purposes and indeed teachers have been left with the challenge of converting teaching materials into mechanisms for assessment (Douglas, 2000, p.ix).

In line with the growing significance of EAP and its development as an area of teaching and learning, as well academic research, the theme of EAP assessment has also been growing in importance (Blue et al., 2000, p.8). With particular regard to the need for Assessment Literacy associated with language testing and more specifically EAP, it should be noted that the use of the term Assessment Literacy has been common amongst researchers in language teaching and general education in recent times. This has emerged through a consensus in the field of language teaching and general pedagogy that there is a need to describe what language teachers need to know about assessment matters (Inbar-Lourie, 2008; Malone, 2011; Stiggins, 1991; Taylor, 2009) whether they are involved in the process of selecting, administering, interpreting, and sharing

results of large-scale tests produced by testing or examination boards, or in producing, marking, analysing, and enhancing in-house or classroom-based assessments (Taylor, 2009, p.24).

Given the researcher's experience as a university language centre director and EAP professional in the UK higher education sector, the context of this study has been particularly influenced by current trends in the British university sector, which have ramifications for professional practice.

International student recruitment has become a fixed priority for revenue generation both for the UK and international university sector (Hyland, 2006). In the UK, due to the governmental capping of home-student recruitment and the unrestricted revenue opportunities presented to British Higher Education Institutions (HEIs) by the recruitment of non EU overseas students, the importance of EAP assessment has grown, as a mechanism for increasing the number of qualified and full-fee paying international students who are eligible to study at university level. This situation has also created extensive business opportunities for large-scale standardised testing systems and strategic openings for individual institutions which can offer alternative in-house programmes in EAP development, such as pre-sessional courses.

As a result of the situation described above, the subject of Assessment Literacy is also considered by the researcher to be particularly worthy of further investigation. A connection can be made between Assessment Literacy in EAP and the social impact of EAP assessment practice (Benesch, 2001; McNamara & Roever, 2006; Shohamy, 2001). For these reasons, ethicality will also be referred to during the course of this thesis, as good practice emerging from language testing research and proponents of Assessment Literacy retains this objective as a reoccurring thread.

1.4 Chapter conclusion

This introductory section of the thesis has described the circumstances and motivational factors which drive this professional-practice-focused research agenda. The significance of Assessment Literacy in the field of EAP has also been introduced so that the importance of research of this nature could be clarified.

The research questions which underpin the enquiry that will be undertaken have been presented. This has highlighted the objective of investigating research-led pedagogical practice in EAP assessment, in order that levels of EAP practitioner Assessment Literacy can be sustained or enhanced in line with identified good practice.

Now that this introductory stage has been completed it is possible to proceed to Chapter 2 in order to explore relevant literature and conceptual frameworks in further depth.

CHAPTER 2 LITERATURE REVIEW

2.1 Chapter introduction

Building on the rationale for research, which has been established in Chapter 1, this section of the thesis will now explore research material which is relevant to the research questions identified. The research agenda which underpins this investigation into EAP teacher Assessment Literacy is supported by a body of literature which spans the fields of Education, Applied Linguistics and EAP assessment. The areas of focus, selected below, have been identified due to their relevance to the research questions used for this thesis:

- Key proponents of Assessment Literacy
- Features of Assessment Literacy and examples of supporting research in Education, Applied Linguistics and EAP
- Guidance resources for the development and enhancement of EAP Assessment Literacy
- Training for teacher Assessment Literacy through general and EAP teacher education
- University-led EAP Assessment Literacy through in-house testing

In line with the particular research questions, as identified in Chapter 1, the content of the Literature Review will explore the nature and source of EAP teacher exposure to research-led assessment practices and assessment-related professional development opportunities, which can be linked to Assessment Literacy.

The issues which emerge from the relevant literature will now be explored in further detail, in accordance with the sources of research as identified above.

2.2 Key proponents of Assessment Literacy

2.2.1 Section introduction

Over the last two decades, a number of experts in the field of education have drawn more explicit attention to the need for educationalists and other stakeholders to develop Assessment Literacy. This situation has also led to a more granular conceptualisation of what Assessment Literacy entails and can be considered useful when seeking to conduct research into the Assessment Literacy of EAP practitioners.

According to Price et al. (2012, p.9) Assessment Literacy can be considered as a gateway or threshold to further learning as it affords individuals not only the necessary skills and knowledge about assessment good practice but facilitates the evaluation of educational situations and decision making regarding which assessment-related skills should be deployed when and for what purpose. This ethos aligns with recent approaches to university education including the teaching of EAP (Alexander et al., 2008, p.309), where there is a movement towards explicit and transparent reference to assessment practices. This approach includes making reference to the assessment criteria used, so that stakeholders, including students, are made fully aware of the intricacies of the assessment process and what is expected of them.

What follows is a review of research and recommendations undertaken and presented by notable experts in the emergent field of Assessment Literacy. This represents a key step towards addressing the research questions of this study by identifying what researchers consider Assessment Literacy to comprise, in terms of skills and practices.

The Material which follows is divided into three main sections, the first two of which present the work of two key campaigners in the field of Assessment Literacy. This is then followed by an additional section presenting the work of a number of other supporters of Assessment Literacy with an emphasis on the significance of Assessment Literacy to foster learning.

2.2.2 The work of Stiggins in the field of Assessment literacy

Stiggins (1991, 1995) is generally accredited with introducing the term Assessment Literacy, the different aspects of which are summarised in Table 2 below :

Table 2 Aspects of assessment literacy, (Stiggins, 1995, p.240):

Assessment Literacy involves:	
1	Approaching assessment with the full knowledge of what is being assessed;
2	Understanding the purpose of assessment;
3	Familiarity with the available means of assessing the achievement in focus;
4	The expertise to generate sound samples of performance;
5	An awareness of what can go wrong and how problems can be prevented before they arise.

Stiggins (1995) directly confronts the dangers which can arise when tests are poorly built and inaccurately interpreted. The problems which stem from poor Assessment Literacy are highlighted by identifying shortfalls in the social norms associated with assessment, these can be summarised as follows:

- Students are not usually encouraged to act as critical consumers of feedback provided by teachers
- Students are not trained to act in their own best interest when it comes to the evaluation of their own achievement
- Teachers are not always open to critical review of their assessments
- Parents are not always skilled in how to understand the meaning of grades or their children's test scores
- Parents are not typically encouraged to ask probing questions regarding their children's assessment results

- School boards and other members of the community do not always understand what constitutes a high-quality assessment or how to recognise faulty or problematic assessments.

Stiggins (ibid) then demonstrates how the situation is worsened or even perpetuated by those who teach or hold roles in educational leadership who typically have very little explicit assessment related training. As a practical reaction to his concern regarding poor levels of Assessment Literacy, three actions are recommended, below in Table 3, which educationalists, who seek to assess, evaluate and take action based on student test results, should take into account:

Table 3 Stiggins' key actions for educationalists engaged in assessment (based on, Stiggins, 1991, p. 536-538)

Action 1: Understanding Assessment Literacy	The development of a good knowledge of the full range of student achievement targets, learning outcomes and assessment methods (including pen and paper assessment; observation and direct personal communications) available. An assessment literate person understands that effective teaching, learning, and assessment are not possible without such knowledge.
Action 2: Differentiating Levels of Literacy	The enhancement of Assessment Literacy in different ways according to the needs of different sectors of society, including: **Level 1 -** (functional literacy) for users of assessment data produced by others such as students, parents, school board members and legislators. This group need to be able to interpret assessment information and know how to relate assessment data to decision making. **Level 2 -** (practical literacy) - individuals who produce and use data, including teachers, managers, teacher trainers and academic advisers. This group needs the functional skills, as mentioned above but also needs to be able to produce assessments which provide useful and trustworthy information which encourages on-going student learning. The skills of this group also include understanding the limitations of different assessment types. **Level 3-** (advanced literacy) for the creators of high stakes tests and individuals who work in the field of educational measurement. These individuals must develop the skills associated with Level 1 and Level 2 but must also develop key skills in standardised test development and evaluation along with an understanding of the challenges presented by large-scale test administration and the maintenance of validity and reliability.
Action 3: Assessment literacy programmes	The need to provide training in different forms, based on the varying needs of different user groups and stakeholders, in order to elucidate the process of educational measurement and its interpretation. This should include : • University programmes • Textbooks • In-service training

In an update to his original crusade for Assessment Literacy Stiggins (1995) identifies certain positive developments which can be considered achievements to have emerged through a heightened awareness for good practice in assessment, these include:

- A societal shift recognising the need to move beyond the ranking of achievement
- A realisation that schools need to help students achieve ambitious achievement targets
- The revision of published educational goals
- The enhancement of performance assessment and the use of new forms of observance and judgement for assessment.
- Increased involvement of teachers in major assessment programmes
- The endorsement of classroom competencies for teachers by educational associations

Nevertheless, crucially, Stiggins (ibid) identified certain persistent barriers to the more comprehensive spread of higher levels of Assessment Literacy, these included:

- Fear of assessment
- Insufficient time for assessment
- Erroneous public perceptions about the general high quality of assessment

In conjunction with the identification of these on-going barriers, the need for bespoke professional development is advocated as the only means of rectifying the situation. Unfortunately, despite the clarity of this advice, given the relative unfamiliarity with the term Assessment Literacy, there still appears to be a great deal of work to do before such recommendations are adopted more widely and integrated into training programmes and CPD offerings for teachers.

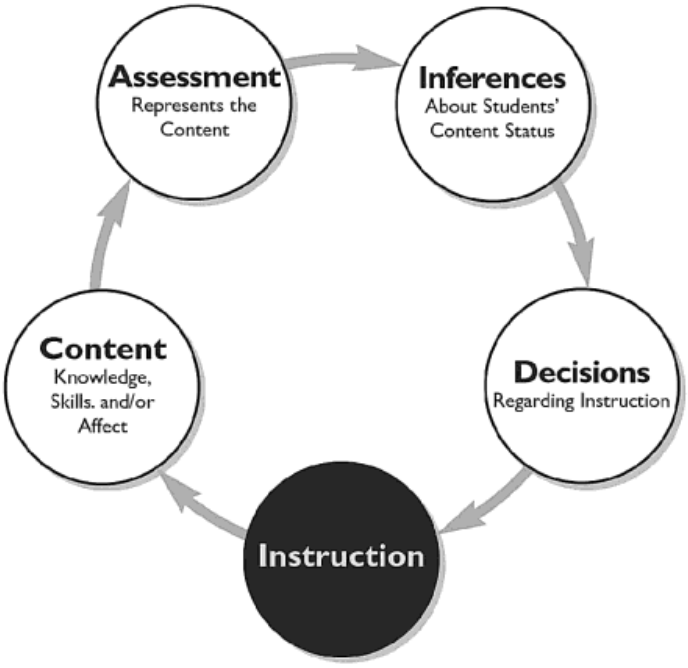
2.2.3 The work of Popham in the field of Assessment Literacy

In his various reports on Assessment Literacy, Popham (2001, 2006, 2009) challenges the quality of high stakes testing, the outcomes of which can have key implications for test-takers' lives, and the negative impact which this is having on students in education.

Popham (2001, p.29) identifies the ideal function of assessment in the instructional process and, in a post-modern approach, which echoes some of the views of Foucault, challenges the inferences drawn from tests whilst inviting the reader to probe further into what a test really demonstrates. Popham (2001, 2006) questions the reliability of inferences drawn from test scores which he believes are inherently flawed. As a means of highlighting the gap in practitioner expertise, Popham (2006) makes an allegory between the medical profession and explains how undesirable it would be if doctors were similarly unskilled and unable to understand patients' medical charts.

As Figure 1 below demonstrates, students' overt performances on sample-based tests can help teachers arrive at inferences about the students' covert knowledge and skills. However, for Popham (2001, p.15), the most crucial consideration in the creation of high-stakes assessments is whether the testing tools used actually assist or obstruct the quality of learning which students experience.

Figure 1 Assessment's ideal role within the instructional process (Popham, 2001, p. 29)



One key quality of Popham's work (2001, 2006, 2009 and 2012) is his promotion of proactivity amongst assessment stakeholder groups.This is evident through the series of action challenges which summons educators to play a role in the process of Assessment Literacy by:

- Offering Assessment Literacy programmes to teachers and administrators
- Establishing autonomous parent action groups
- Delivering assessment related briefing sessions for educational policy makers
- Introducing a public information campaign regarding high-stakes tests
- Conducting rigorous security monitored reviews of items from high-stakes tests
- Implementing defensible evaluative schemes at local level
- Demanding the installation of a more educationally appropriate state-wide test

In Popham's (2006) account it is explained that, in order to enhance Assessment Literacy it is also necessary to campaign amongst the following groups of society who are all players and stakeholders in the act assessment:

- Educators
- Parents
- Everyday citizens
- Students themselves

In addition to the negative effect which poor testing can have on test takers, Popham (2006) also advocates the need for Assessment Literacy from the perspective of the educator. This view suggests that most educators are unfamiliar with how large scale tests are constructed and yet their professionalism is frequently measured against students' performance on such tools. With regard to EAP Assessment Literacy a comparison could be made with pre-sessional courses which aim to take students beyond training for a standardised test such as IELTS, despite the lack of access to detailed grade descriptors for such tests, which make the design of such follow-on courses very complex.

Popham (ibid) attempts to focus on and demystify the challenge of becoming assessment literate by referring to a relatively small set of principles in an attempt to banish concerns that assessment requires the understanding of a series of 'obtuse notions'. Reference is also made to the fact that courses in measurement are not sufficiently connected to the realities of contemporary education and that individuals who follow such educational measurement programmes often report that the syllabus is 'quantitatively intimidating and instructionally irrelevant'.

Popham (2009, p.11) goes on to explain that Assessment Literacy is a commodity needed by teachers for their own on-going professional benefit and for the educational well-being of their students, as teachers will continue to need to operate in an educational context where test-elicited data plays an important role. However, as Popham (ibid) does not feel that pre-service courses provide a sufficient level of Assessment Literacy for newly qualified teachers, in the same manner as Stiggins (1995) suggests, the need for professional development programmes, including those which promote Assessment Literacy, is still very much a requirement.

The key series of principles which Popham identifies as fundamental to the acquisition of Assessment Literacy (in Popham, 2006) are presented in further detail below (Popham, 2009, p.8-10) in the context of requirements for Assessment Literacy courses:

1. The fundamental function of educational assessment is the collection of evidence from which inferences can be made about students' skills, knowledge and affect, namely maintaining candidate interest and motivation in test taking (Weir, 2005). This also incorporates washback and ethics as we collect evidence to understand traits that we cannot see
2. Reliability of educational assessments, especially the three forms in which consistency evidence is reported (stability, alternate form and internal consistency)
3. Validity of evidence supporting test-based inferences about student skill
4. How to identify and eliminate assessment bias that offends or unfairly penalises test-takers because of personal characteristics such as race, or gender
5. The need to be conversant with the processes involved in creating and honing test items

6. Scoring of student responses using well-formed rubrics (criteria) and avoidance of loose judgemental procedures. The need to know how to create and use rubrics, namely scoring guides so that constructed responses can be accurately appraised
7. How to generate and mark a suitable variety of assessment strategies, given the range of items and assessment types students are presented with
8. The understanding of the learning potential of formative testing as demonstrated by research (Biggs & Tang, 2011, p.64; Knight, 1995)
9. How to collect and interpret evidence of students' attitudes, interests and values so that learning and assessment can be effectively pitched and attuned.
10. Interpreting students' performances on large scale standardised achievement and aptitude assessments as these are of relevance and interest to students and parents including what is meant by a scale or score
11. Becoming conversant with assessment procedures most suitable for subgroups of society including people with disabilities and people from other cultures
12. How to appropriately and inappropriately prepare students for high-stakes tests
13. How to determine the appropriateness of an assessment mechanism. Simply because a tool has been adopted doesn't mean it is suitably well-constructed or well-interpreted

With the aim of promoting the proactive acquisition of skills in Assessment Literacy amongst educators, Popham (2012b) has also created a series of 15 booklets which attempts to guide educators on a need-to-know basis through the key aspects of developing skills in assessment. These resources offer a step-by- step approach which can be applied to a wide range of contexts in a manner which is accessible, even from the lay perspective and attempts to establish a voice which replicates a peer-to-peer conversation. Whilst these resources certainly have a niche in that

they appeal to educators who may be overwhelmed by other more technical volumes, anyone wishing to gain a deeper level of expertise would undoubtedly need to read more widely.

2.2.4 Recent examples of Assessment Literacy in action

In addition to the work undertaken by prominent researchers, such as those mentioned in the previous section, who have explained the importance of Assessment Literacy in the wider field of education, a number of examples of developments associated with Assessment Literacy in practice highlight how the concept is being embraced across the sector. This includes schools and universities, both in terms of EAP and education on a more general level, in addition to the emergence of approaches which prioritise stakeholders' needs. Three examples will be provided below:

2.2.4.1 Assessment Literacy with Assessment for Learning

Assessment for Learning positions student assessment at the heart of an integrated approach to student learning and stresses the fact that high quality and appropriate assessment is a prerequisite for better student learning (Knight, 1995).

According to the former Department for Children, Schools and Families (DCSF, 2008), 'Assessment for Learning is an important way of improving students' attainment and is founded on the premise that learners will improve most if they are able to understand the objectives behind their learning, where they are positioned in relation to this objective and how they can achieve the goals which are being targeted.' Consequently, Assessment for Learning is arguably essential to any effective form of education, where assessment is present or required. Assessment for Learning

can be defined as the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there (Assessment Reform Group, 2002). As a result, the promotion of Assessment for Learning and its empowerment for stakeholders, including test-takers and teachers who create tests, has certain key shared goals with campaigns for Assessment Literacy.

Popham (2009, p.11) highlights the power of Assessment for Learning in a manner which demonstrates how Assessment Literacy can be applied by stating that, 'when classroom assessments are conceived as assessments for learning, rather than assessments of learning, students will learn better what their teacher wants them to learn.'

A more detailed definition of Assessment for Learning (DCSF, 2008), which highlights how this application of Assessment Literacy can be harnessed, states that good assessment for learning makes:

- an accurate assessment – knowing what the standards are, judging pupils' work correctly, and making accurate assessments linked to National Curriculum levels;
- a fair assessment – knowing the methods used are valid
- a reliable assessment – ensuring that judgements are consistent and based on a range of evidence;
- a useful assessment – identifying barriers to pupil progress and using that information to plan and discuss the next steps in learning.
- a focused assessment – identifying areas of a child's learning where there are blocks to progression, which might, for example, benefit from the attention of one-to-one tuition
- for continuity of assessment, enabling better transfer between years and schools

So popular are movements such as Assessment for Learning that some educational activists such as Black (2003) go as far as to as to recommend the rejection of emphasis on examination results and reference to league tables in favour of an alternative approach which supports Assessment for Learning (Biggs & Tang, 2011, p.64; Knight, 1995). Nevertheless, in the true context of modern day society where examination results continue to play a pivotal role, the requirements of stakeholders may perhaps be better serviced by educationalists such as Gottlieb and Nguyen (2007) who have worked to develop a technique which aims to support the achievement of designated educational outcomes, which may be fixed by national curricular, whilst simultaneously promoting actual progress in transferable learning.

2.2.4.2 ASKe and empowerment of test-takers through Assessment Literacy

Members of the Assessment Knowledge Standards Exchange (ASKe, 2012), a Centre for Excellence in Teaching and Learning (CETL) based at Oxford Brookes University, have been investigating assessment practices in Higher Education (HE) over the last two decades with a view to promoting the importance of Assessment Literacy, with a particular emphasis on student involvement. Whilst the impact on stakeholders, particularly students and test takers, is always central to the promotion of Assessment Literacy, as acknowledged by Stiggins (1991; 1995) and Popham (2001), Price et al. (2012) from ASKe offer a particularly prominent focus on the importance of fostering the Assessment Literacy of students and their understanding of the mechanisms of assessment. This view of the critical agenda for Assessment Literacy emphasises the broader definition of assessment and refers not only to measurement of achievement but also to the accessibility and interpretability of feedback as well as the giving of support and development of student learning. This approach towards formative assessment shares attributes with Assessment for Learning (Biggs & Tang, 2011, p.64; Knight, 1995), as mentioned earlier.

The stance adopted by Price et al. (2012) acknowledges the fact that the more students understand what is expected of them in the assessment process the more likely they will be to meet the demands of the assessment mechanisms they experience. It is also recognised that assessment is central to the student experience and that student satisfaction ratings for assessment and feedback in HE remain less than optimal. Price et al. (2012, p.10) describe Assessment Literacy in terms of knowledge, skills and competencies. In this view, Assessment Literacy encompasses an appreciation of the purposes and processes of assessment which allows deep engagement with assessment standards and the making of choices about which skill or area of knowledge to apply. This facilitates an understanding of situations when aspects of assessment are appropriate or inappropriate. This definition of Assessment Literacy extends to include:

- An appreciation of assessment's relationship to learning
- A conceptual understanding of assessment (i.e. understanding of the basic principles of valid assessment and feedback practice, including terminology used)
- Understanding of the nature, meaning and level of assessment criteria and standards
- Skills in self- and peer-assessment
- Familiarity with technical approaches to assessment (e.g. pertinent assessment and feedback skills, techniques and methods, including their purpose and efficacy)
- Possession of the intellectual ability to select and apply appropriate approaches and techniques to assessed tasks (which skill to use when and for which task)

2.2.4.3 A Constructivist approach to fostering Assessment Literacy

Inbar-Lourie (2008) considers the core skills required in order to undertake assessment in the contemporary educational context, in this account the importance of multiple forms of assessment and the contextual evaluation of language skills is recognised. An environment is promoted which seeks to replicate the true roles of language users and which adopts a social reconstructivist paradigm. Inbar-Lourie (2008) notes that, whilst some language testing courses maintain a focus on test design and proficiency assessment, others go as far as to include coverage of the social implications of language testing. This approach aligns with the aim to avoid any negative implications of misrepresentative data, as supported by critical language testers, such as McNamara (2001), McNamara and Roever (2006) and Shohamy (1998; 2001).

Inbar-Lourie (2008) makes reference to the gap which appears to have arisen between psychometric approaches to language testing and measurement and more interpretive Vygoskian methods. Vygoskian approaches consider learning and knowledge to be a dynamic socially co-constructed phenomenon with language playing both a key role in the shared understanding and the fostering of student learning. In this Vygoskian paradigm, where learning is jointly constructed and students are viewed as key participants in the learning process, the formative view of assessment for learning is seen as paramount. The Vygoskian approach to assessment (Daniels, 1993) promotes a dynamic method of assessment which allows for the exploration of what is referred to as the zone of proximal development and thus potentially takes assessment away from the confines of typical mono-dimensional assessment types and promotes social interaction with peers in order to allow for the collaborative construction of meaning which may not be possible in isolation.

The challenges facing EAP teachers in promoting a more holistic and interpretivist view of assessment are highlighted by Lynch & Shaw (2005, in Inbar-Lourie, 2008) where the growing need for measurement and evidence of minimum proficiency attainment is seen to promote a more positivist testing paradigm. This phenomenon is seen as a result of the kind of data and test-based information that contemporary society and its institutions demand.

According to Inbar-Lourie (2008), the development of Assessment Literacy requires the attainment of a set of competences, some of which are practical and others that are theoretical. With this in mind, the components of Assessment Literacy are described as:

- Having the ability to ask critical questions about the assessment purpose
- Interpreting the fitness of the testing tool being used
- Considering the conditions under which assessment takes place
- Taking and understanding actions based on test results

According to Inbar-Lourie (2008), in order to foster the development of Assessment Literacy a constructivist approach is recommended which, if interpreted in the contexts of this thesis, suggests the need for collaboration between EAP practitioners and assessment specialists.

Three major considerations are identified which require the skills which collectively define Assessment Literacy and thus should drive training courses, these considerations are:

- Why?- The rationale for assessment
- What?-The trait or construct being assessed
- How?-The nature of the assessment process

In Inbar-Lourie's (2008) opinion, courses in Assessment Literacy need to take into account the organic and complex nature of language testing and its varied contexts, as language assessment will be ineffective if it is unrepresentatively limited in form or breadth.

According to Inbar-Lourie (2008), Assessment Literacy should reflect social roles of assessment and theory about how both language and knowledge are best acquired. The manner and approach should be suitably flexible in accordance with the requirements of the different stakeholders involved. In this approach, there should also be a Continuing Professional Development (CPD) approach where practitioners continuously and regularly update their knowledge and skills of assessment.

2.2.4.4 The BALEAP Working Party on Testing

With a specific focus on EAP, a number of key developments have also emerged in recent years from the BALEAP Working Party on Testing (BWPT) which has investigated assessment practices at various stages associated with EAP in university education in the UK (Schmitt, 2012, p.2) including:

- Pre-entry screening
- End of course assessment
- In-sessional screening

Documents produced and disseminated by this group of EAP experts (BALEAP, 2012a; Schmitt, 2012) have provided important insights and resulted in key directives which are relevant and accessible to the wider range of EAP assessment stakeholders. These materials aim to assist all test users in acquiring a level of understanding concerning the relationship between EAP test

purposes, test format and the meaning of test scores so that realistic and fair standards can be applied. Test users are then presented with a series of key questions which should be considered when evaluating a particular assessment as a measure of proficiency. Assistance is also provided in the following key areas:

- Setting and minimum scores for entry or progression purposes
- Considering (Common European Framework of Reference) CEFR requirements and institutional needs
- Determining the length of pre-sessional programmes
- Providing a series of test digests which give an overview and evaluation of tests commonly accepted for University entry purposes

As well as the reviews provided of standardised tests such as IELTS, TOEFL and PTE Academic, the group also warns of the need to give further attention to in-house testing systems, given the extended and ambitious aims of most EAP training courses and assessments and the limitations of the normal duration of pre-sessional courses.

In addition to the contribution which BWPT makes to the enhancement of EAP Assessment Literacy, the need for an on-going collaborative approach through assessment communities is highlighted in the following statement (Schmitt, 2012):

> 'High stakes is happening in classrooms – we need to work together more to support EAP practitioners developing greater Assessment Literacy and practical skills.'

2.2.5 Section conclusion

This element of the Literature Review has scrutinised the key aspects of Assessment Literacy as promoted by some of the most prominent and most acclaimed advocates of Assessment Literacy.

As a result, this has enabled the identification of key factors and a range of stakeholder standpoints that need to be taken into consideration when promoting the development of Assessment Literacy. Through this process, a more detailed definition of the term Assessment Literacy has emerged, which is a key aspect of addressing the agenda required by the research questions for this thesis. This definition can now be used in the next section of the review as a framework for the purpose of a broader investigation of supporting literature in Education, Applied Linguistics and EAP.

2.3 Features of Assessment Literacy and examples of supporting research in Education, Applied Linguistics and EAP

2.3.1 Section introduction

The research questions which steer this research project require an investigation of testing and assessment practices which are considered to comprise Assessment Literacy. Consequently, the key aspects of Assessment Literacy, as described by Stiggins (1995), in Table 2 on page 28, will be used as a means of organising a series of relevant examples from available literature. Stiggins' approach has been selected for this purpose as it provides a series of broad categories; however it is notable that many of the key factors associated with Assessment Literacy, as described by Popham, (2009), are also relevant within Stiggins' definition.

Although the examples provided are by no means exhaustive, the aim of this exercise is to provide illustrations which highlight ways of developing and maintaining a more systematic approach to assessment in a manner which aligns with the principles of Assessment Literacy in contexts relevant to EAP.

2.3.2 *Approaching assessment with the full knowledge of what is being assessed*

In this area of Assessment Literacy, Stiggins (1991, 1995) identifies the need for educationalists to develop Assessment Literacy in order to understand more comprehensively what is being tested. The examples which have been selected to demonstrate resources or relevant research for practitioners working towards this aim include research into construct validity and the outcomes-based approach to assessment.

2.3.2.1 Understanding what is being assessed- the role of construct validity

In the past, some tests and assessments have developed using less principled approaches than those which have emerged from more recent research into assessment good practice (Fry et al., 2003, p.44). Support for a more scientific paradigm for testing and Assessment Literacy can be accredited to research into psychological testing and the enquiry conducted by Cronbach (1990, p.192) and Messick (in Linn et al., 1989, p.16-17). Work conducted by these researchers into measurement has in turn assisted in the development of the academy of language testing as a result of investigations into reliability and validity. Consequently, familiarity with construct validity represents a key aspect of Assessment Literacy of practitioners seeking to develop or use EAP assessments. The connection between EAP assessment and construct validity is further strengthened by Blue et al. (2000, p.26-27) who share the belief that construct validity is an overarching form of validity which is crucial to EAP assessment and testing.

Researchers have made wide reference to the importance of construct validity in the context of tests and assessments over the course of the history of language testing research. It was perhaps

Lado (1961) who was the first contemporary author on the topic of modern language assessment who wrote about validity. Construct validity is of key relevance to Assessment Literacy (Moss et al., 2006, p.116) as an understanding of this area is critical to the building of quality tests and assessments and the avoidance of the negative influence of poorly operationalised constructs (Fulcher, 1999, p.226; Messick, 1989, p.226).

In some cases, attempts to explain construct validity have appeared to complicate the matter by describing in detail a range of sub forms of validity such as content validity or face-validity (Davies, 1968), rather than the unified and thicker description as supported by others (Henning, 1987; Hughes, 2003; Messick, 1989). On the other hand, other explanations could be seen as an underrepresentation of validity due to a failure to take into account the social and political implications of testing and the dangers of misinterpreting results (McNamara & Roever, 2006; Shohamy, 2001; Stobart, 2008).

Messick (1989) is recognised as changing the way in which validity is defined and approached through his description of construct validity as a unitary concept with various facets (Fulcher, 2010, p19). See Table 4 below:

Table 4 Messick's classic validity framework (1989, p.20)

	TEST INTERPRETATION	TEST USE
EVIDENTIAL BASIS	Construct validity	Construct validity + Relevance/utility
CONSEQUENTIAL BASIS	Value implications	Social consequences

In Stiggins' terms (1995), construct validity can be seen to assist as it provides a fuller knowledge of what is actually being assessed. McNamara (1996, p.18) also warns of the dangers associated with restricted views of validity in EAP assessment which do not embrace the fuller breadth of construct validity. Certain questions which are posed by Messick (1989, p.16) assist with probing further into what is being tested as part of a process of validation:

- Does the content of the assessment method adequately match the content and objectives of the instructional program?
- Is the student responding and reacting to our assessment methods and enquiries in expected ways?
- Does the assessment evidence relate in an expected way to external behaviours?
- Does the assessment method allow us to investigate student differences due to various educational conventions?
- Is the assessment evidence sensitive to the social consequences regarding its use?

Wigdor and Garner's definition of construct validity (1982, p.62) aligns with the understanding of construct validity as a feature of Assessment Literacy as it is associated with deeper knowledge about what is being assessed in the process of testing; it is explained that construct validity 'is a scientific dialogue about the degree to which an inference that a test measures an underlying trait or hypothesised construct is supported by logical analysis and empirical evidence'. This view is supported by recommendations made by Henning (1988) who calls for the empirical validation of EAP tests based on operational test data.

It should be noted, however, that the outright nature of some validity claims, such as those made by Cronbach and Meehl (1955) have since been challenged. Fulcher and Davidson (2007, p.11) clarify that the connection between validity theory and evidence is not always clear due to the fact

that theories develop and evidence of validity can continue to be researched. As a result, assertions about the extent of any test's validity should not be treated as permanent or impermeable. Indeed, one key point to make is that no test in itself should be described as valid, but rather it is the inferences drawn from the evaluation of test data which can be considered to be to, a greater or lesser extent, valid (Popham, 2012c, p.9).

In order to see how different approaches to construct validity may have informed Assessment Literacy in the field of language assessment, it is useful at this point to consider the different principles which have guided researchers in this area.

Chapelle (1998) defines a construct as 'a meaningful interpretation of observed behaviour' and acknowledges that the difficulty in defining constructs lies in 'hypothesizing the source of performance consistency'(Chapelle, 1998, p.34). With this in mind, Chapelle describes two different perspectives on test response consistency, the 'trait theorists' who attribute consistencies in test response to test-taker characteristics and 'behaviourists' who link consistencies to contextual factors, see Table 5 below:

Table 5 Trait principles and behaviourist principles (Chapelle, 1998, p.42)

	Trait principles	Behaviourist principles
Meaning of performance	A sign of underlying characteristics	A sample of performance across similar contexts
Construct definition	Implicit characteristics must be specified independent of context	Implicit characteristics cannot be specified. Context must be specified
Test construction	Random sampling of content allows generalisability across all contexts Contextual influences are considered irrelevant Minimise these by presenting and eliciting language out of context	Careful selection of content allows generalisability across similar contexts Contextual influences are considered relevant. Maximise these by presenting and eliciting language in relevant contexts
Validity justification	Compare performance on different test methods to distinguish method error from the trait variance	Compare the test context to the context of interest to identify similarities and differences

Notably, the approaches listed above represent different models of inquiry associated with construct validity which in turn represent key aspects of Assessment Literacy. The polarised perspectives also highlight how views relating to different aspects of Assessment Literacy may differ according to the philosophy of the practitioner concerned. Contemporary language testing researchers now accept that it is important to understand how traits and contexts interrelate and as a result, this has led to the emergence of a third perspective known as interactionalism which encompasses both the language knowledge of trait theory and the contextual factors of behaviourism (Bachman, 1990, p.84). In one sense, this interactionalism and the understanding of the broader context through a consideration of both trait and behaviourist theory follows the same track as a quest for Assessment Literacy which takes into account the broader spectrum of skills and capabilities required of language testing and assessment practitioners.

As demonstrated in this section, an understanding of the concept of construct validity constitutes a key element of any aspect of Assessment Literacy which involves understanding what is being assessed. As construct validation involves investigating the extent to which an assessment

measures the aptitude, or skill which it intends to assess (Heywood, 2000) it could be argued that Assessment Literacy cannot fully exist without a firm grasp of the notion.

2.3.2.2 Considerations for domain sampling in EAP Assessment Literacy

Kane (2012, p.41-42) explains the challenges presented to educators by the aspect of Assessment Literacy associated with domain sampling, or collecting representative examples of target language usage from a real world context. As Kane (ibid) explains, identifying the range of performances to include in a target domain and deciding on the emphasis to be given to different kinds of performance tends to be complicated when the domain needs to be used in the development of an assessment. In Kane's view, certain key considerations are required, when deciding how to approach the sampling process. These considerations are summarised below:

- When drawing performances from a broadly defined and diverse target domain, the sampling variability is likely to be quite large. This can lead to difficulties when considering the extent to which samples can be used to generalise about wider populations and can also be a source of measurement error.
- When samples are drawn from individuals who are active in the target domain, it needs to be taken into account that performance will be affected by knowledge and skills which are not associated with language. If this is not considered then it can lead to construct-irrelevant variance (Messick, 1989).

In relation to EAP, the considerations as highlighted above demonstrate that a key aspect of an approach to sampling and consideration of domain relates to the practitioner's stance on the matter of specificity. With this in mind, it is useful to consider debates within EAP which focus on the

general or specific purposes dilemma. Robinson (1980, p.13) and Widdowson (1983 p.6-11) did make an effort to describe the great challenge which EAP teachers are confronted with in attempting to successfully and purposefully analyse and describe the spectrum of potential different academic discourses and genres within an EAP classroom, however notably, no practical solution was proposed.

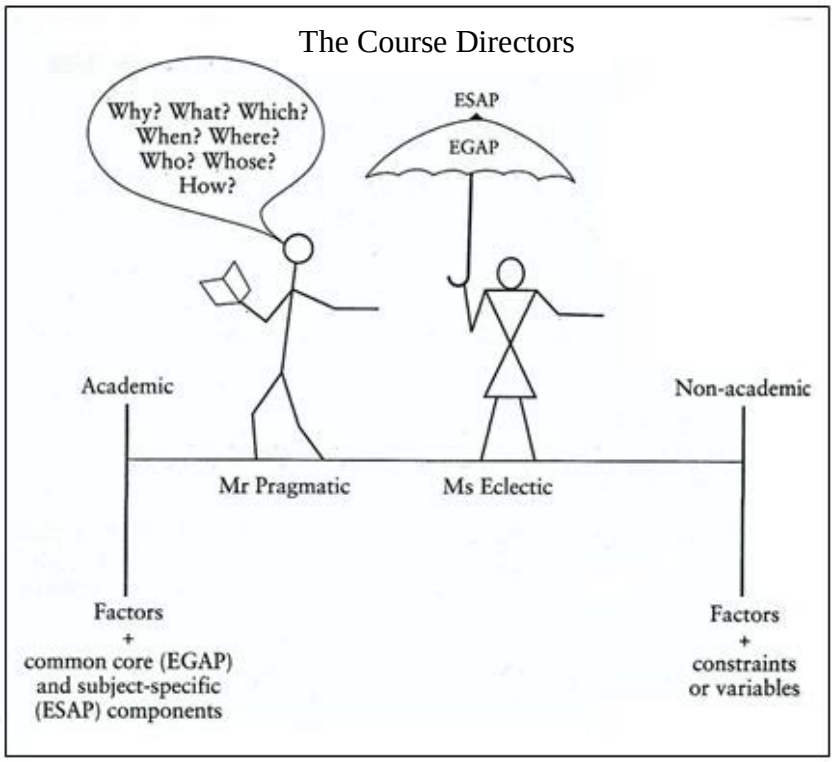
On the other hand, in the development of ELTS, it can be seen that Carroll (1981) did seek to use an ESP approach to testing (Clapham, 1996) when test versions were developed on the basis of Munby's (1978) communication needs processor model. Furthermore, Swales (1985), did also offer a number of real-world case studies through his important collection of articles which presents a series of different ways in which the various challenges of the science-and-technology-related EAP/ESP classroom have been met.

Since the 1970s and 80s controversies have continued to emerge associated with the choices and challenges brought about by favouring either process or product. Consequently, debates surrounding English for General Academic Purposes (EGAP) and English for Specific Academic Purposes (ESAP) began to thrive (Jordan, 1997; Selinker et al., 1981; Spack, 1988; Hutchinson & Waters in Swales, 1985; Widdowson, 1983). Seminal articles by Spack (1988) and Dudley-Evans and St John (1998) highlight continuing disputes surrounding what should constitute the domain for EAP and whether there should be emphasis on ESAP or EGAP involving either a subject-specific approach or a focus on a common core of performance which some believe is shared across academic sub-domains.

In order to highlight the complex challenges facing EAP course directors, including domain sampling, Jordan (1997,p.66) illustrates, in Figure 2 below, the multi-layered series of

considerations which have to be deliberated during the process of creating and delivering of EAP programmes. These same considerations have a key impact on the validity and reliability of tests and assessments, although notably this understanding was not specifically referred to in the original diagram below:

Figure 2 The course design balancing act (Jordan, 1997, p.66)



It is clear that choices made by assessors in the process of developing or promoting Assessment Literacy can have a crucial impact on the sampling process and ultimately the EAP test or assessment which is created. However, it should be noted when considering the role of domain sampling in Assessment Literacy for EAP test development, language testing researchers interested in EAP assessment (Fulcher, 1999; Widdowson, 1983) have criticised the traditional focus and debate in EAP surrounding content and specificity as they believe it has caused distraction from what is important in the process of assessment, namely construct validity which ensures that decisions made on the basis of test scores are trustworthy, because the inferences drawn from students' grades are reliable and valid.

One practical example of EAP sampling in practice which can act as a useful case study for Assessment Literacy in EAP relates to the work undertaken by Weir (1983). Weir's research interest aligned with that of Carroll (in Alderson & Hughes, 1981, p.67) who believed that content validity could be achieved in the testing of EAP through analysing test takers and their needs and using this information to determine the content of tests. Central to Weir's research was a behavioural analysis of students' communicative situation in order to determine more accurately what language and communication skills were involved. The main means of establishing content validity was through authentic sampling and representation of the disciplinary domain (Fulcher & Davidson, 2007, p.6).

After isolating the disciplines and the respective demands of degree programmes most frequently studied by international students at the time, Weir wished to explore the viability of harnessing the results of his study in order to construct a test specification (Waters, 1996, p.42-44). In other words, in order to pursue the concept of a test which was based on communicative reality, Weir aimed to build a test which effectively and more precisely reflected target University communication activities and circumstances under which they were normally performed (Weir, 1988, in Hughes, 1988, p.46). One major criticism of Weir's focus on the communicative situation is that its preoccupation with content, authenticity and sampling does not address the deeper questions about ability and competence which are now acknowledged as key to construct validity (Fulcher, 1999, p.223). Furthermore, the difficulty of accurately identifying what constitutes 'authentic situations' was also acknowledged.

In addition, although authenticity is not usually cited as a critical quality of language testing, Bachman and Palmer (1996, p.23-24) argue for an important link between authenticity and

construct validity by determining that authenticity determines the level to which score inferences can be considered representative of the target language use domain. However, it should be noted that Lewkowicz (2000) claims that it is difficult to ascertain the way in which test takers perceive authenticity and the definition of authenticity may be variable according to different stakeholders. Weir's research is relevant to the field of Assessment Literacy as it represents a design for a proficiency test based on sampling conducted through empirical research into student needs. The central view that EAP tests should flow as naturally from needs analysis as the EAP course itself is however challenged by Fulcher (1999, p.221) who argues that concerns for authentic content should not obscure preoccupation with the main question of how valid inferences are drawn from test scores.

Weir's work will be considered in further depth later in this review, in Section 2.6, when the development of the TEEP test will be considered in terms of its contribution to our understanding of Assessment Literacy.

2.3.2.3 Understanding what is being assessed- the outcomes based model

If a key feature of Assessment Literacy involves understanding what is being assessed, then research associated with the outcomes-based approach to education and assessment can also be considered to offer an important resource to those seeking to cultivate or encourage Assessment Literacy in the field of EAP. In the view supported by Driscoll and Wood (2007, p.5), by identifying learning outcomes, or stated expectations of what someone will have learned and how they will be assessed, student centred learning can be promoted along with the improvement of curriculum and pedagogy. Such good practice forms a key feature of many contemporary EAP courses which now commonly seek to specify learning outcomes in order to provide transparent

assessment goals for students and to adhere to requirements for accreditation as set by professional bodies such as BALEAP (Alexander et al., 2008, p.309). Some of the promising practices of good assessment which exemplify the outcomes based approach are also cited by Driscoll and Wood (2007, p.4):

- Assessment information should be provided to students in advance of instruction
- Students should be able to direct their learning efforts to clear expectations
- Student progress and completion of learning outcomes are determined by achievement of learning outcomes

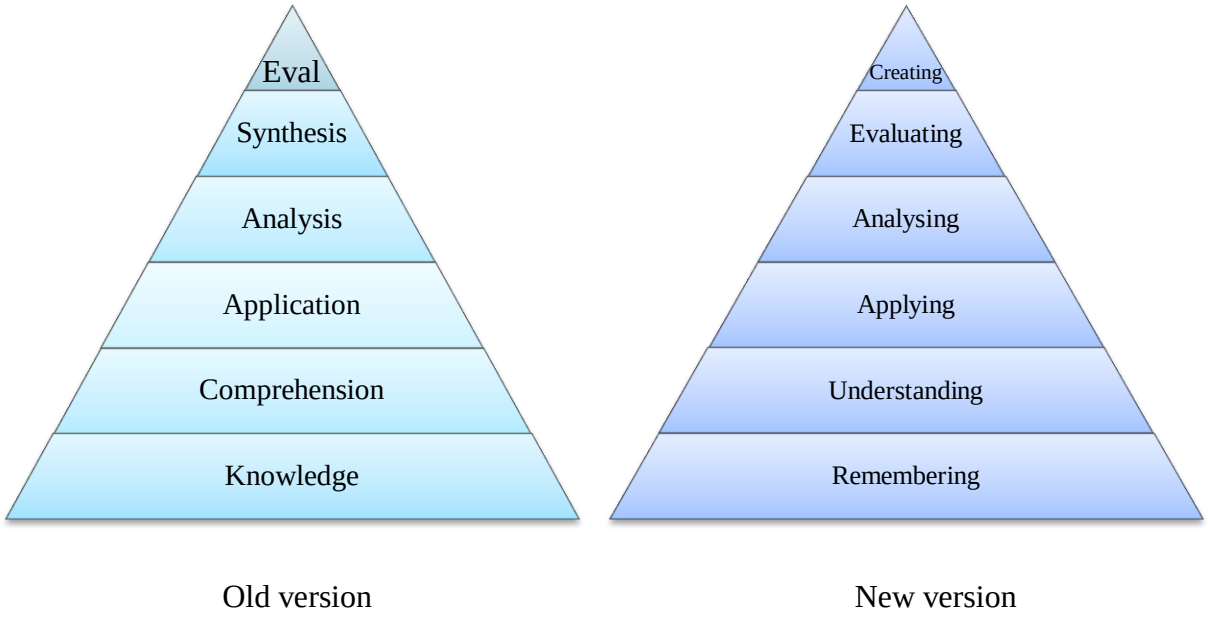
Driscoll and Wood (ibid) also make the point that learning outcomes need to play a dynamic role in the structuring and development of curriculum.

Bloom's contribution to the outcomes-based approach to education (starting with his work in Bloom,1956) continues to have a key influence on the field of Assessment Literacy for educationalists involved in teaching and learning, as it encourages reflective examination of educational practice and supports in the process of identifying suitable teaching, learning and assessment strategies (Fry et al., 1999, p.40). In this way it aligns with Stiggins' recognition of the importance of knowing what is being assessed.

The focus of Bloom's particular educational research was to classify the processes involved in learning (Forehand, 2005). The result, Bloom's Taxonomy, was a framework designed to organise learning into different stages and attempt to add transparency to learning process (Fry et al., 1999, p.33). Since its design more than 50 years ago, and despite later revisions, such as Anderson and Krathwohl (2001, p.67-68), the original taxonomy continues to be used widely in the process of

determining and measuring learning outcomes in education, including at tertiary level in specification documents for modules and specifications (Fry et al., 1999, p.33). Indeed, with specific reference to the EAP context, the importance of the processes as identified in Bloom's Taxonomy is highlighted by Green (2007, p.48) who explains that in academic writing the stages of the taxonomy are used to both build and assess student understanding.

Figure 3 Bloom's taxonomy (old version) and (new version) as updated by Anderson & Krathwohl (1994, p.67-68)



According to Stiggins (1995), Bloom's Taxonomy presents an excellent foundation for teaching and assessing reasoning; this has key relevance to Assessment Literacy as the skills which are required involve being prepared to define, teach, and assess the kinds of capabilities which Bloom identified as common in a wide range of educational contexts. A useful example of how Bloom is still referred to in the use of grading criteria for assessment purposes in higher education; an excerpt from a document produced by Northumbria University (2011) is presented below in Figure 4:

Figure 4 Assessment guidelines from Northumbria University (2011) referring to Bloom's taxonomy

…guidance that has been issued by the University has been based in Bloom's taxonomy, providing a set of verbs relating to different categories of learning outcome. While the categories do not map exactly onto levels, it is expected that higher levels will involve more learning outcomes and assessment tasks requiring the more complex activities:

Lower level	Knowledge	define, repeat, record, list, recall, name, relate, underline
	Comprehension	translate, restate, discuss, describe, recognise, explain, express, identify, locate, report, review, tell
	Application	interpret, apply, employ, use, demonstrate, dramatise, practice, illustrate, operate, schedule, sketch
	Analysis	distinguish, analyse, differentiate, appraise, calculate, experiment, test, compare, contrast, criticise, diagram, inspect, debate, question, relate, solve, examine, categorise
	Synthesis	compose, plan, propose, design, formulate, arrange, assemble, collect, construct, create, set up, organise, manage, prepare
Higher level	Evaluation	judge, appraise, evaluate, rate, compare, revise, assess, estimate, predict

With reference to the situation facing EAP contemporary students, Alexander et al. (2008, p.309) describe the content of University courses which has been made more explicit through specifying content in terms of learning outcomes which emphasise what students should be able to do rather than what is in the syllabus. The impact of the outcomes based approach on EAP is further highlighted by Sheldon (2004, p.152) who reiterates that the measurement of EAP skills, in particular academic writing, is best undertaken through an outcomes-based approach as it prepares students to do and not just to know.

2.3.3 Understanding the function and purpose of assessment and how this relates to EAP

As Stiggins (1995) explained, Assessment Literacy involves knowledge about the purpose of testing and assessment. This section will focus on material which indicates the importance of understanding test purpose and function, in order to clarify the significance of this aspect of Assessment Literacy.

As clarified by Berry (2008, p.13), assessment is used for a range of different functions which can be categorised in a series of differing ways. In broad terms, these can be divided into two areas, namely for drawing inferences related to the performance of individuals or the effectiveness of the system, perhaps linked most closely to summative assessment and secondly for the improvement of learning and the provision of feedback, also described as formative assessment (Biggs & Tang, 2011, p.141). Table 6 below then extends these basic functions by describing a series of additional purposes.

Table 6 Functions of assessment (Berry, p.13)

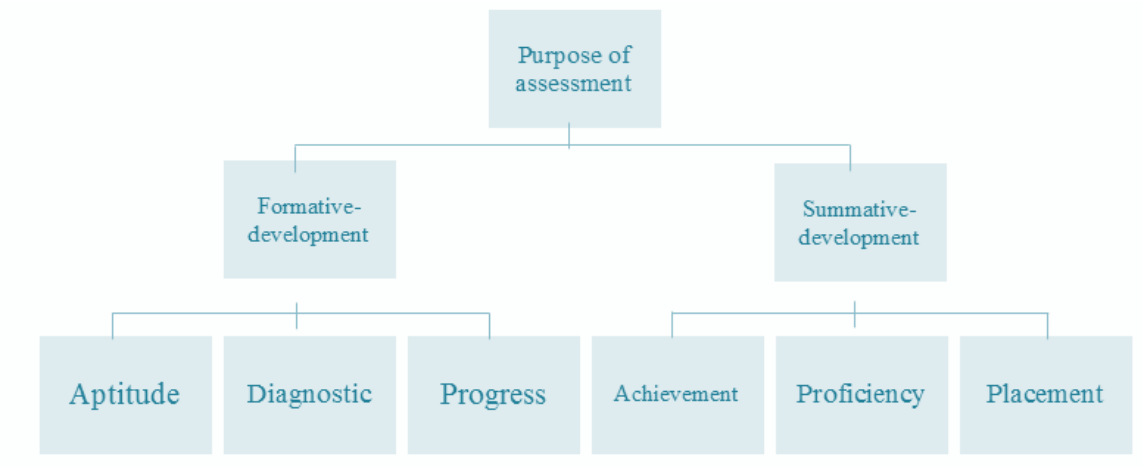
Assessment Function	Description
Selection and Placement	Test or examination is used to determine who will be selected to university, or is used to place students into school bands
Accountability	Assessments are used to determine if students have achieved learning outcomes appropriate for their grade level, and are used to judge the effectiveness of teachers, schools etc. in helping learners to achieve the intended learning outcomes
Diagnosis	Assessment is used to identify underlying sources of learning difficulties.
Support of learning	Assessment (especially at the classroom level) are used to monitor the progress of learning; to provide learners with feedback on their learning, to help them improve; to assist teachers in identifying changes to be made in their teaching; to enhance student motivation and confidence by demonstrating progress.

Alexander et al. (2008, p.307) describe the complexities of trying to separate the functions of tests into formative and summative categories, as some summative EAP tests may also have a formative function whilst a test designed for EAP proficiency measurement may also yield diagnostic information. Crucially, however, it is identified that it is important to be clear about test uses so that measures can be taken to ensure that tests are appropriate for the selected purposes. An example is given of the use of a proficiency test for a short intensive pre-sessional EAP course which may not be appropriate if there is not sufficient time for students' overall level of proficiency to improve.

When displayed in this manner, it becomes clear that the task of understanding test purpose is a complex and multi-faceted aspect of Assessment Literacy.

Arguably one of the key language-assessment texts providing a user-friendly overview of the field, which is accessible to the lay language tester is provided by Hughes (2003, p.11-25) who describes a range of different tests and their respective purposes, with the needs of practising language teacher in mind. Figure 5 below, provided by Alexander et al. (2008), summarises the main purposes of tests whilst relating them to EAP contexts and also stresses the key considerations associated with formative and summative situations.

Figure 5 The purpose of assessment (Alexander et al., 2008, p.304)



In addition, Davies (1990, p.20) concisely describes the distinction between purposes of a test and its uses. In Davies' view, the primary purposes of language tests are selection, feedback, evaluation and research. On the other hand, the uses of a test are commonly referred to in terms of achievement, proficiency, aptitude and diagnosis. This description is extended by Henning (1987, p.4-8) in, Table 7 below which describes different test types and includes reference to varying approaches and options in testing situations:

Table 7 Henning's test types (1987, p.4-8)

Objective tests:	Subjective tests:
Comparing test responses with a set of acceptable responses such as in a key.	Scoring based on judgement and expertise.
Direct tests:	Indirect tests:
Rating of language in a relatively real and situation such as an interview test for spoken language assessment.	More contrived mechanisms for the assessment of language such as multiple choice questions.
Discrete-point tests	Integrative tests
Tests designed to measure knowledge of performance in restricted areas, such as verb usage, in isolation.	Tests designed to measure a greater range of language abilities simultaneously.
Criterion referenced tests	Norm referenced tests
Tests which measure achievement on the basis of the degree of mastery of a pre-specified domain/ criteria or set of learning outcomes.	Tests which evaluate performance of test takers on the basis of the relative achievement of other test takers.
Speed tests	Power tests
Items are so easy that every person may be able to get every answer correct given enough time; however sufficient time is not provided so both knowledge and speed of performance is assessed.	Sufficient time is allowed for all candidates to finish however such difficult items are included that not all tests takers are expected to answer every item correctly or fully.

Fulcher,(2006) and Fulcher and Davidson (2009) have explained the importance of test purpose in a way which is clearly understandable to the lay assessor or educationalist seeking to make clear

the importance of this aspect of Assessment Literacy. In this example architecture is used as a as a metaphor for test development:

‘When architects begin to design a building, they must have a very clear idea of its purpose. If a client wishes to open a supermarket there is little point in designing a neoclassical residential town house. If the requirement is for a space to repair cars, the architect would not design a restaurant and if I wished to build a country retreat where I could get away from the city… I would be rather upset if the architect produced plans for a block of flats. Similarly the materials needed for the construction of these buildings would be different, and the cost of the buildings would vary accordingly.’(Fulcher, 2010, p.95).

2.3.4 Familiarity with the available means of assessing the achievement in focus;

The works of Davidson and Lynch (2002), Fulcher and Davidson (2007, 2009) and Fulcher (Fulcher, 2006, 2010, 2012a, 2012b) can be considered to have made a key contribution to the series of resources available for the development of Assessment Literacy including guidance materials available to EAP teachers. One particular key contribution can be considered to be in the provision of guidance associated with processes involved in assessment and how these can be considered in a more systematic fashion. In this manner, research into test design and the use of specifications is particularly relevant as an aspect of Assessment Literacy.

2.3.4.1 Test design and use of specifications- tools for Assessment Literacy

The Test Design Framework (Fulcher, 2010, p.94), see Figure 6 below, provides a useful guidance tool for EAP teachers in the process of navigating the various stages of test design (Davidson & Lynch, 2002; Fulcher & Davidson, 2007); this might include stages such as construct definition, item writing, prototyping and piloting. With this aim in mind, it can be argued that the guiding function of this model can be used as a frame of reference irrespective of the resource or amount of time available to the practising EAP teacher.

Figure 6 The test design framework

Supporting and encouraging the use of test specifications to facilitate a more scientific and reliable approach to test design is also a key aspect of the work undertaken by Davidson and Lynch (2002), Fulcher and Davidson (2007, 2009) and Fulcher (Fulcher, 2006, 2010, 2012a, 2012b). The use of test specifications could also be considered an important guidance tool for any assessment

literate test developer. According to Davidson and Lynch (2002, p.3) the test specification is the primary tool for language test development. In the view of these researchers, by using a specification, even the lay language tester can enhance their assessment practice.

Ruch (1924) provides an early example of an approach to using test specifications which can be harnessed to facilitate the creation of objective examinations in a manner which allows for flexibility according to the context of particular testing requirements or institutional constraints. In language testing, the term 'specification' is used to describe a generative blueprint document through which alternative versions of a particular assessment or assessment task can be created (Davidson & Lynch, 2002, p.4). As Davidson and Lynch (2002, p.3) put it, "a specification is an efficient generative recipe for a test that fosters dialogue and discovery at a higher abstract level than achieved by analysis of a simple item or task."

Alderson (2000, p.168) and Alderson et al (1995) also consider the use of test specifications to be of vital importance to test design and development, in conjunction with the various stakeholders involved in the process of test design and use. These stakeholders include test writers, test validators, teachers and of course, test takers. In addition, research into psychological test construction conducted by Anstey (1966, in Davies, 1990, p.12) demonstrates that the initial stage of test development requires extensive planning. Davies (1990, p.12) explains the relevance of psychological testing to the language testing model and supports the use of specifications in the initial planning stage. The need for the use of specifications at the initial stages of language test construction is also advocated by Hughes (2003, p.26).

With the specific context of EAP in mind Alderson et al. (1995, p.21-25) provide a framework for the design of a test specification which is a useful starting point for EAP practitioners seeking to

consider the available means of constructing a template. Further details regarding the development of ESP and test specifications and the process of needs analysis is also provided Alderson (1988).

A prototype test specification was developed by Popham (1971) and this has been further adapted by Davidson and Lynch (2002, p.14) See Figure 7 below:

Figure 7 Test specification format (Popham, 1971 in Davidson & Lynch, 2002, p.14)

Specification Number: Provide a short index number
Title of Specification: A short title should be given that generally characterizes each spec. The title is a good way to outline skills across several specifications.
Related Specification(s), if any: List the numbers and/or titles of specs related to this one, if any. For example, in a reading test separate detailed specifications would be given for the passage and for each item.

(1) *General Description* (GD): A brief general statement of the behavior to be tested. The GD is very similar to the core of a learning objective. The purpose of testing this skill may also be stated in the GD. The wording of this does not need to follow strict instructional objective guidelines.

(2) *Prompt Attributes* (PA): A complete and detailed description of what the student will encounter.

(3) *Response Attributes* (RA): A complete and detailed description of the way in which the student will provide the answer; that is, a complete and detailed description of what the student will do in response to the prompt and what will constitute a failure or success. There are two basic types of RAs:
a. Selected Response (note that the choices must be randomly rearranged later in test development): Clear and detailed descriptions of each choice in a multiple-choice format.
b. Constructed Response: A clear and detailed description of the type of response the student will perform, including the criteria for evaluating or rating the response.

(4) *Sample Item* (SI): An illustrative item or task that reflects this specification, that is, the sort of item or task this specification should generate.

(5) *Specification Supplement* (SS): A detailed explanation of any additional information needed to construct items for a given spec. In grammar tests, for example, it is often necessary to specify the precise grammar forms tested. In a vocabulary specification, a list of testable words might be given. A reading specification might list in its supplement the textbooks from which reading test passages may be drawn.

Bachman and Palmer (1996, p.177) clarify that a test specification can be used for evaluation purposes in order to consider the:

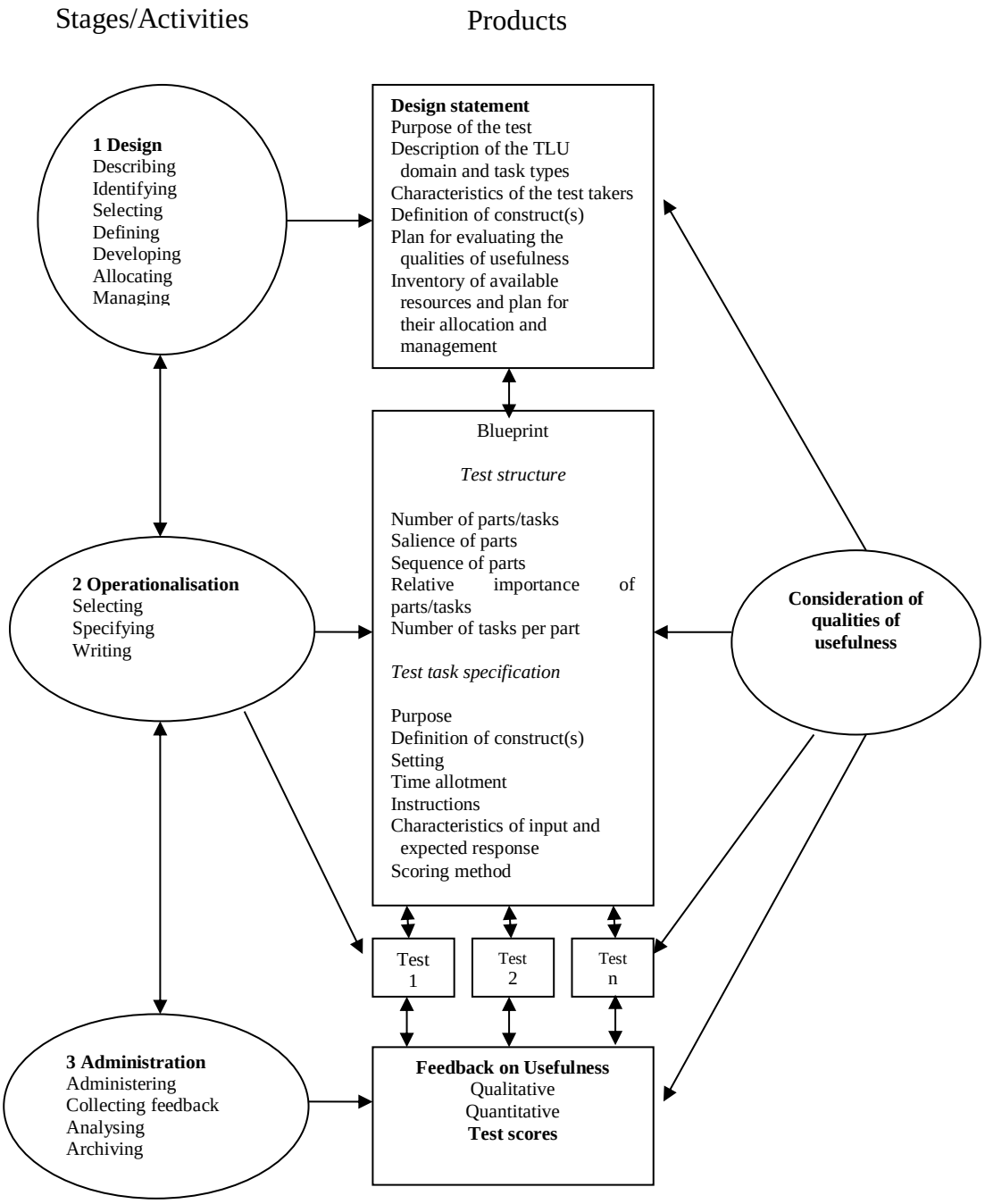
- intentions of the test developer
- correspondence between the test and the specification used to develop it
- relationship or authenticity of tasks in the test compared to the target language domain

Bachman and Palmer (1996, p.87) also describe the test development process in terms of design, operationalisation and administration which is illustrated in Figure 8 on page 66. Although Alderson (2000, p.168-169) recognises the merits in this approach, an alternative process is suggested below:

- Identifying test purpose
- Developing test specifications
- Guidelines for and training of item/task writers and moderation of their products
- Pre-testing, analysis of results and revision of test
- Training examiners and administrators
- Monitoring examiner reliability
- Reporting scores and setting pass marks
- Test validation
- Post-test reports
- Developing and improving tests

This model claims to position the specification more centrally, thus avoiding extensive overlap across component stages.

Figure 8 Overview of test development (Bachman & Palmer, 1996, p.76)



With the aim of explaining how language test specifications can be of widespread benefit, Fulcher and Davidson (2007, p.60) make an allegory with the field of engineering and state that 'tests should be built to the best of an organisation's ability… and when the test needs refitting, the organisation should refit it.'

In order for test specifications to be used to their best advantage, additional advice is provided by Fulcher and Davidson (2007, p.61) who advise of the importance of allowing specifications to develop through discussion, trialling and piloting in 'an creative, organic, consensus driven, iterative process'. This view is reinforced by Jafarpur (2003, p.72) who recommends the formation of testing-related groups in order that team contributions can be coordinated and recorded so that deviations from test specifications can be avoided. Davidson and Lynch (2002) are also keen supporters of the collaborative and team-based approach to test creation, although, Davies (1990, p.13) suggests that 'whether item writing is done by one person or by a working party is incidental'. Jafarpur (2003) clearly disagrees and recognises that without team contribution the isolated test or item writer risks adding him or herself as a test facet. Bachman (2000, p.72) extends this view and recommends 'the development of standards of practice and mechanisms for their implementation and enforcement'. However, Jafarpur (2003, p.72) warns against compelling item writers to use very rigid guidelines and specifications which may result in what is described as test cloning, namely the design of tests which lack the contribution of writer ingenuity.

Evidently, in reviewing different approaches to the application of specifications, those seeking to promote or develop Assessment Literacy may wish to bear in mind that it is the use of a specification per se which is considered most crucial rather than the particular manner in which it is employed, as this will vary according to the context and purpose of testing and assessment (Davidson & Lynch, 2002, p.20).

2.3.4.2 Marking, grading and giving feedback

With key aspects of Assessment Literacy in focus, a familiarity with the available means of assessment can also be interpreted in terms of marking, grading and the process of giving feedback. With this in mind, a number of key texts provide useful guidance relating to skills required by language testers and offer the EAP practitioner a source of practical advice on creating and using scoring and feedback procedures.

Fulcher and Davidson (2007, p.91) explain the importance of scoring by explaining that the test score is the link between the evidence we elicit from the task on the one hand and the construct and domain on the other. Fulcher and Davidson (2007 p.91-114) also clarify the significance of a principled approach to the process of scoring, given the different nature of reasons and requirements for testing. In some situations, where the objective of testing is predominantly formative in nature, the key objective may be to provide feedback to students, whereas in other situations there may be higher stakes where grades are used to draw further inferences or for gatekeeping purposes.

Bachman and Palmer (1996, p.51-52) describe two key processes which should be undertaken in order to attribute scores to students' performance in tests or assessments, the first of these includes identifying the number of items which have been successfully answered or completed. This approach involves:

- specifying criteria in order to determine what constitutes successful completion
- deciding whether responses will be scored as either correct or incorrect or with varying levels of correctness

Alternatively scoring can be undertaken according to levels of language ability. This criterion based approach involves:

- identifying rating scales in order to assess the language
- considering how many stages of ability are to be incorporated into the different scales

This criterion based approach links back to outcomes-based planning, as outlined by Fry et al. (2003, p.26-41) and is discussed in further detail on page 54.

As Hughes (2003, p.94) explains, This latter approach to scoring can be described as analytic and commonly involves the award of a separate score for each of a number of aspects of a task, such as lexis, grammar and coherence.

Another approach to scoring, known as holistic scoring exists which is a more impressionistic scoring involving the assignment of a single score based on overall impression. Whilst this approach has the advantage of being very rapid, it makes use of less tangible or scientific reference points and is often based on intuitive shared understandings within teams. It could therefore be challenging to apply in a principled manner in many situations.

A further consideration highlighted by Blue et al. (2000, p.13) relates to different ways in which marking is undertaken across different departments. Blue et al. (ibid, p.13) make reference to a research project involving both international and home Electrical Engineering students whose

written reports were marked by both EAP and Engineering staff using different criteria and returning different results. This collection of articles also makes reference to the predictive validity of test scores obtained by students taking pre-sessional courses in EAP.

Hughes (2003 p. 94-100) advises that any scale that is used for scoring should initially be calibrated. This process should involve gathering representative test result samples which represent a model of ability which spans the range of test taker performance. The assessment team should then examine the samples and award a mark using the scale. This process can then act as a form of training and can also be used as a means of identifying how useful or reliable the scale itself is. In addition to the importance of scale calibration, Hughes stresses the importance of rater training and outlines a three step training approach.

With regard to the interpretation of test scores, Bachman and Palmer (1996, p. 226), like Fulcher and Davidson (2007, p.91) also remind us that score interpretations have two components which are:

- Inferences about the components of language ability that have been rated and
- Inferences about the Target Language Use (TLU) domain to which these inferences generalise

As a result of these aspects it is recommended that score reports provided to test takers and other stakeholders should include a clear definition of the construct as well as a description for the type of tasks and scoring methods involved. From the test taker's perspective, this also complies with principles of assessment for learning (Biggs & Tang, 2011, p.64; Knight, 1995) and allows for formative benefits from assessment feedback even in situations where summative outcomes determine access or progression.

If a fuller description of what the test scores mean is not available to test consumers, both the test taker and other individuals or institutions who are required to consult a grade, then a situation can arise where certain scores attract meanings which may not be wholly accurate. As Fulcher and Davidson (2007, p.92) suggest, this happens with testing organisations such as ETS (2011), providers of the TOEFL examination along with The British Council and IDP, providers of IELTS (2008, 2012), who put the onus on institutions to interpret scores without the provision of a great deal of background information.

2.3.4.3 Reference materials for the purpose of statistical analysis

This section refers to resources which may be of use to educationalists seeking to acquire proficiency or foster interest amongst other language testers in the process of developing key skills in statistical analyses. It should not be confused with the section of this thesis which analyses the data collected in this research project. In many cases the skills honed by EAP teachers are different to those of the statistician, indeed some language teaching professionals may even have chosen their career to avoid the scientific disciplines (Bachman, 2004, p.ix), consequently, given the close association of high-stakes testing with measurement and aspirations of precision, it is necessary for EAP language Assessment Literacy to involve development of at least some skills in using statistics. Key works which are relevant, and accessible to contemporary EAP practitioners are produced by experts in this field, such as Bachman (2004), Bachman and Kunnan (2005) and Larson-Hall (2010). These resources offer an invaluable transferable toolkit for the purpose of applying statistical procedures specifically to language testing situations, relevant to EAP; they also provide electronic resources linked to software packages such as SPSS and Excel so that users can put exercises into practice.

Benefits of the use of statistics are described in terms of the manner in which they facilitate analysis of complex numerical data and the identification of trends (Woods et al., 1986, p.1). In addition, in line with a multidimensional view of validity (Messick, 1989, p.20), Gorard (2001, p.5) states that qualitative research which is completed without any concern for quantitative support risks compromising its value.

One of the most difficult choices presented to the researcher involves identifying the most appropriate statistical analysis tools to use (Brown,1988, p.113), as reliance on the incorrect technique will produce inaccurate or misleading results.

Inferential statistics are usually believed be more useful in terms of generalisation than descriptive alternatives (Gorard, 2001, p.150); this is because they provide a better gauge of how the analysis of a sample may relate to larger groups of people (Brown,1988, p.115). Such data enables the researcher to step beyond descriptive analysis of the particular sample collected by generalising findings to the wider population (Dörnyei, 2003, p.115; Larson-Hall, 2010, p.45).

Tables 8 (A-C) and Tables 9 (A-D), starting on page 74, present and explain some of the statistical procedures which are described in a number of the key texts and which should be considered by EAP practitioners seeking to develop this aspect of Assessment Literacy. Methods suggested in the tables below can be used for purposes such as to:

- investigate and display the distribution of marks within a set of test scores to see how grades in a test are spread
- consider how grades are grouped and to identify average or central tendencies within a group of scores

- examine the variability of distribution of test scores and the difference between the highest and lowest scores from a test
- reveal relationships between different variables, such as different sections within a test
- explore reliability of test items so that unreliable items can be discarded and replaced
- assess the reliability of item sets, sections or test versions or the reliability of raters
- inspect the validity of constructs within tests

Tables 8 A-C Descriptive statistical analysis procedures and resources for EAP language testers

Objective A
To investigate the spread or distribution of scores or grades (Bachman, 2004, p.41-54; Dörnyei, 2007, p.208; Larson-Hall, 2010, p.245)
Purpose/further explanation
distribution is typically used in order to: • Describe the shape and characteristics of a score dispersal • Identify outliers or anomalous scores • As a precursor to using descriptive statistics of grouping **Often scores are arranged or displayed in a(n):** • Ordered list • Tabular form • Histogram/bar chart, • Box and whisker plot • Frequency polygons Kurtosis (in a frequency polygon) The level of peakedness in symmetrical distributions: Leptokurtic – highly peaked Platykurtic – relatively flat Mesokurtic – middle peaked (typical of normal distribution) Skewness (in a frequency polygon) Distributions which show a form of asymmetry Positively skewed – longer tail at upper end & peak at lower end Negatively skewed – longer tail at lower end & peak at upper end Bimodal distribution – two peaks

Objective B
To examine how grades or scores are grouped (Bachman, 2004, p.55-62; Kirk, 2008, p.68-73)
Purpose/further explanation
Descriptive statistics of grouping are indicators of central tendency, often expressed as a Number, or percentage and displayed individually or collectively in tabular form or marked on a frequency polygon. **Mode** The mode (the most frequently occurring score) of different groups is useful for describing demographic information such as age, sex and level of education, to understand characteristics of test takers. **Median** The median splits the data into two equal sets. And can also be expressed as the same figures as the second quartile (Q2), This is a point in the distribution above which exactly half of the scores occur and below which the other half occur. **Mean** Mean of a test score is often interpreted as an indicator of test difficulty or how a given group of students has performed. It is calculated as the sum of the scores gained on the test divided by the number of test takers to which the scores pertain.

Objective C
To describe the variability or distribution of test scores (Bachman, 2004, p.63-67; Kirk, 2008, p.106)
Purpose/further explanation
Range: The range shows the difference between the highest and lowest scores. It can be calculated by subtracting the lowest score from the highest score. It does however include extreme values which can skew perceptions of dispersion. The range can be expressed as at interval such as 5-75 (5, lowest and 75 highest) or an interval width such as 70 (for range above). **Interquartile range:** Inter-quartile range is the difference between the upper quartile or 75th percentile (Q3), and the lower quartile, 25th percentile (Q1). The formula is therefore (Q3 - Q1). Effectively this procedure finds the difference between the smallest and largest observations within the middle half of the observations. When compared to the range, the inter quartile range is a more effective measure of central tendency as it focuses on the spread of the middle half of the values. It is based on the principle that, for any symmetrical (not skewed) distribution, half of its values will be found one semi-interquartile range either side of the median (Q2). If the data set contains outliers, the range alone can appear very large however will not represent the data properly. In contrast, the inter-quartile range is not influenced by outlying data. **Semi interquartile range/quartile deviation:** The semi-interquartile range is the interquartile range divided by two (Q3 - Q1) ÷ 2. It is a measure of variability which is frequently used for data sets which are not normally distributed. As the semi-interquartile range spans one half the distance between (Q1) and Q3) it is even less affected by extreme scores than the interquartile range. When data distribution is very skewed and normally distributed, it is more useful to use the semi-interquartile range to describe dispersion along with the median. **Standard deviation:** The standard deviation shows how much, on average test scores vary or deviate from the mean. In a situation where there is a form of normal distribution, if cases are fairly tightly clustered together and the bell-shaped curve is steep, the standard deviation is small. When the examples are distributed further apart and there is a flatter bell curve, there will be a comparably larger standard deviation. When the mean is used as a measure of average, the standard deviation is usually used as the measure of dispersion.

Table 9 A-D Inferential statistical analysis procedures and resources for EAP language testers

Objective A
To investigate relationships between two different entities (Bachman, 2004, p.80-85; Dörnyei, 2007, p.223-225; Gorard, 2001, p.167-170)
Procedure(s)/Further explanation
Correlation coefficients are statistics calculated from data which summarise the strength and direction of the relationship between two variables. Correlation coefficients range between negative one (-1.00) and positive one (+1.00). Positive coefficients indicate direct relationships while negative coefficients indicate inverse relationships. The larger the coefficient, positive or negative then the stronger the relationship. A correlation near to 0 suggests a weak relationship and a correlation near to one (either positive or negative) shows a strong relationship. **Pearson product-moment correlation coefficient:** This procedure requires the relationship between the two variables to be linear, both variables should constitute interval scales and both should be normally distributed. **Spearman rank correlation:** This procedure is used to investigate the relationship between two variables without interval measures or normal distribution.

Objective B
To investigate the reliability of different test items (Bachman, 2004, p.129-131; Fulcher & Davidson, 2007, p.87-88, p.103-104)
Procedure(s)/Further explanation
Item difficulty index **Item discrimination index** **Item-total test score correlations** The above item statistics can be used to: • give feedback to test takers and teachers and test developers • help identify problems with items which can help with control characteristics of score distribution • identify weaknesses for correction, to increase the internal consistency reliability of a test • ascertain why items are not operating satisfactorily

Objective C
To test hypotheses between different groups or variables; to investigate internal reliability of test scores; to test reliability of grading procedures (Bachman, 2004, p.159-174; Dörnyei, 2007, p.215-219; Fulcher & Davidson, 2007, p.105; Larson-Hall, 2010, p.241-243, p.268-269).
Procedure(s)/Further explanation
The procedures below are used to compare the scores of different groups who have taken the same test or the scores from a single group that has taken different tests: **T-test:** used for two groups **ANOVA (Analysis of Variance):** for more than two groups These procedures are used to identify differences between two or more different groups of test score or class performance based on different variables such as teaching methodology. These analyses can help answer questions such as; if the level or quality of one variable is modified then will this also affect the level of the other variable? Independent (un-matched) sample T-tests are for research where it is necessary to compare the scores of groups or classes that are independent of each one another. For example, Class A and Class B. Paired-sample T-tests are used in research where it is necessary to compare two sets of scores (two variables) collected from the same group or class, for example Class A's scores on two different tests. This type of T-test is also used when the same individuals are tested more than once, such as via a pre-course test and a post-course test. The procedures below are used to examine internal consistency within a test and reliability of items and sections: • Split half estimates (Guttman / Spearman-Brown) Norm Referenced (NR) • Test takers' results on a test are split into two halves to examine reliability across two parallel sets • Estimates based on item variances (Cronbach's Alpha) (NR) • The scores obtained on individual parallel test items are examined for variance • Test-retest reliability estimates- also known as stability test (NR) • To give the same test to the same group of test takers twice with a time delay between administrations and to analyse the variance in results. The procedures below are used investigate reliability of procedures associated with marking and awarding grades; • Rater (Marker) reliability estimates (NR) • This can be used to determine inter-rater or intra-rater reliability • Phi coefficient dependability Criterion Referenced (CR) • This procedure is used to measure the dependability of a test score as a measure of mastery in a skill • Agreement indices(CR) These are used to measure reliability of different classifications related to mastery or non- mastery

Objective D
To investigating test validity (Bachman, 2004, p.257-293; Fulcher & Davidson, 2007, p.184)
Procedure(s)/Further explanation
Use of expert raters: Raters are skilled practitioners in a particular field whose expert judgement is used to quantify suitability of test content often using rating scales **Collection of verbal protocols:** Analysis of test taker experience given verbally after a test has been taken **Exploratory Factor Analysis:** Analysis of patterns in test scores of test takers who have taken a number of different test versions **Multi-trait multi method correlation:** The observance of patterns across a set of abilities measured in different ways via a number of different tests **Confirmatory Factor Analysis:** Patterns of factor loadings specified and expected after initial research observation which are subsequently tested against a data set

Through consideration of the resources above, which present the practical application of statistical analysis procedures, the EAP practitioner can become familiar with a means of providing quantitative evidence in order to support test use (Bachman, 2004, p.x). Given the high stakes nature of EAP, this aspect of Assessment Literacy is crucial as it assists in enabling us to demonstrate whether or not tests are reliable and therefore can be used to mitigate the risk associated with unfair decision making and the impact which this can have on individuals.

2.3.5 The expertise to generate sound samples of performance

Reference to the ability to gather sound samples of performance relates to reliability in the assessment process. Whilst reliability is a condition of construct validity and therefore test usefulness, reliability alone is not indicative of construct validity (Bachman & Palmer, 1996, p.23). Given this distinction and Stiggins' (1995) identification of this area as an aspect of Assessment Literacy, it is useful to describe further this facet of assessment in relation to EAP Assessment Literacy.

Reliability, often described in terms of consistency of measurement, is an essential measurement quality of any test Blue et al. (2000, p.35). In other words, a test score which is reliable will be consistent across different characteristics of the testing situation (Bachman & Palmer, 1996, p.19-20). Reliability also provides a major justification for using scores for the basis of making decisions. The need for EAP tests, which are often high-stakes and used for gate-keeping purposes, to be open to scrutiny and criticism of such factors, which would include reliability is identified by Blue et al. (2000, p.35). Whilst many Higher Education professionals are aware of the admissions function to which results for tests such as IELTS and TOEFL are applied, in-house testing systems developed and used within universities also play a key gate-keeping role. On the completion of pre-sessional EAP courses students either meet or fail to meet linguistic thresholds and are often turned away from their intended courses of study, even though they have already spent some weeks studying within the same institution on language improvement programmes. This therefore suggests a critical need for EAP practitioner competence in identifying features of a reliable or unreliable measure. Such concerns align with Popham's (2001, 2006) scepticism regarding the reliability of inferences drawn from test scores and his concerns regarding the lack of expertise pertaining to many educationalists. Similarly, this situation has led Popham to call for proactivity in the training of educators and other stakeholders in this aspect of Assessment Literacy. It is also why accreditation schemes, such as that provided by BALEAP, are key in promoting high standards of assessment in EAP.

In relation to the EAP assessment context, Hyland (2006, p.100) clarifies that assessment can be considered reliable if it measures consistently in terms of the same student on different occasions, and the same task across different raters. With these factors in mind, Hughes (2003, p.44) provides a useful description of procedures which can be employed for the enhancement of reliability. This

series of steps can be used as a practical tool by EAP practitioners seeking to develop an understanding of the key factors in reliability and thus can serve as a tool to enhance or promote Assessment Literacy.

Table 10 Procedures to enhance test reliability, inspired by Hughes (2003, p.44)

Reliability can be enhanced through:	
1	collecting a sufficiently high number of samples
2	excluding in terms which do not discriminate well between weaker and stronger students
3	not allowing unnecessary freedom in question choice in order to allows more direct comparison between samples
3	writing unambiguous test tasks and instructions
5	ensuring that tests are laid out clearly and legibly
6	familiarising students with the test format
7	providing uniform and non-distracting test conditions
8	using items which allow objective scoring
9	providing a detailed scoring key
10	training raters (agree acceptable responses and scores)
11	identifying candidates by number
12	using multiple independent scoring methods (moderation/double marking)

The need for training in rating, for practitioners involved in the EAP assessment process is further highlighted by Hyland (2006, p.302) who states that assessments should be reliable and all assessors should agree on the criteria and how they will be applied; this requires rater training with the aim of consistent scoring in mind.

With reference to the features of test usefulness in EAP, Alexander et al. (2008, p.314-316) remind practitioners of the fact that all tests, including those focusing on EAP, make compromises and that language tests cannot measure language behaviour in authentic situations but rather in a restricted test situation. The compromises made by large scale standardised testing organisations associated with EAP are also referred to by Alexander at al. (ibid) and EAP practitioners are warned that practicality often dominates over all other factors of test usefulness. With the inherent unreliability of large-scale tests in mind BALEAP (2012a, p.3) advises those reviewing EAP related tests to make themselves aware of the range of scores at which a test discriminates most reliably, in terms of giving an accurate indication of students' proficiency.

Returning to the point made at the start of this section regarding the fact that reliability does not indicate validity; in recent years, the EAP sector has also raised concerns regarding IELTS which are a useful reference point for EAP Assessment Literacy. Due to certain writing tasks used in IELTS which are deemed by some practitioners to be unrepresentative of University writing (Moore & Morton, 2005), it has been suggested that there has been a trade-off between validity and reliability with reliability being inappropriately prioritised by IELTS test designers (Wigglesworth & Elder, 1996, in Moore & Morton, 2005).

Another factor relating to reliability of EAP assessment which is relevant for EAP Assessment Literacy, involves the negative backwash of university administrator familiarity with standardised test scoring formats. An example is provided by an account of practices at Lancaster University (Banerjee & Wall, 2006) which explains how the reporting of EAP grades using a system based on the band-scores used by IELTS was abandoned in favour of writing a profile report for each student due to concerns relating to reliability. Whilst other University admissions colleagues may have appreciated the face validity of scores which appeared similar to IELTS these were ultimately deemed unsuitable by EAP practitioners for the following reasons:

- The official rating scales for speaking and writing were not available
- Estimating reading and listening abilities was problematic, since this depended on inferences about students' 'inner processes' rather than analyses of products such as essays or oral presentations, and
- None of the teaching team had been trained as IELTS examiners.

2.3.6 An awareness of what can go wrong and how problems can be prevented before they arise

Research into the ethics of language testing draws on the methodology of critical social theorists (Bachman, 2000; Lynch, 2001; McNamara, 2001; McNamara & Roever, 2006; Shohamy, 1998, 2001) and builds on Messick's holistic view of validity (Messick, 1989, p.20), see Table 4 on page 46, which includes 'test consequences' and 'social consequences'. Messick recognised that the impact of the test on the stakeholders concerned needed to be taken into account when evaluating the validity of inferences drawn from test scores.

2.3.6.1 Ethicality and social implications- key considerations in EAP Assessment Literacy

Spolsky (1981) may be considered to be one of the first researchers to refer to the ethicality of test use and to discuss the political purposes for which language testing is sometimes used. In Spolsky's view (1981, p.20) language tests should be given a health warning similar to those given to dangerous drugs or chemicals. This is particularly significant in the context of EAP, given the important gate-keeping function which it performs, institutionally, nationally and internationally (Flowerdew & Peacock, 2001, p.192) and the effect that poor assessment or assessment-related decisions can have on people's lives. In particular, this includes larger scale standardised tests such as TOEFL which is a clear example of a high-stakes test with social effects (Spolsky, 2012) given the gate-keeping function which it provides in facilitating or withholding access to University study. Indeed, such is the concern regarding test ethicality that some practitioners (Lynch, 1997; Lynch & Shaw, 2005) have designed alternative approaches to language testing in an attempt to avoid the inherent unethical practices which they believe accompany the power imbalance present in traditional testing systems.

Research undertaken by Shohamy (1998, p.331-332) shows that there is growing interest in the role played by language testing in society. Consequently, topics such as test ethicality and bias are now being discussed in research, publications and conferences. Shohamy (1998, p.332) advocates a critical approach to language testing which acknowledges that the act of testing is not neutral. Shohamy (2001) probes the socio-political agendas and objectives which language testing practices involve. Similarly, Bachman (1990, p.279) supports this view when he refers to the fact that 'tests are not developed and used in a value-free psychometric test-tube; they are virtually always intended to serve the needs of an educational system or of society at large.'

As advocated by McNamara & Roever (2006, p.8) testers need to engage in debate on the consequential application of their tests and need to reflect on test usage after the point of operationalisation, in addition to a customary focus on the test development stages. An example relevant to EAP is provided by Shohamy (2001, p.102) who refers to the impact of an English test in the Middle East due to its usage for university entrance and the anxiety which was caused through only minor changes, given the high-stakes associated with success.

Errors linked to the uses to which tests are put can be linked directly with damaging sociological implications and harmful repercussions. The work of Shohamy (1998, 2001) and McNamara & Roever (2006) are particularly relevant in this respect, as they warn of the power of tests and the potential harm which can be incurred. In addition, these risks also resonate with Foucault's postmodern concerns (Foucault, 1977; Foucault & Gordon, 1980) connected to the interplay between information-seeking and power relations.

With more specific reference to the context of EAP assessment and the need for both EAP practitioners and other stakeholders, including students, to be aware of the societal implications of

EAP assessment, Benesch (2001, p.60) encourages the challenging of conventions associated with EAP through a concern regarding power relations and social justice. Benesch believes that current conditions under which EAP is taught and assessed should be interrogated and probed in the interest of equity and democratic participation both in and outside educational institutions. The ultimate aim of Benesch's work is to assist students in performing well in their academic courses (Benesch 2001, p.xvi).

Concerns of this nature are further highlighted through contemporary accounts of problems with powerful tests such as that provided by Burton (2012). Burton describes the pressures felt by both students and staff as a result of the strict rules and regulations along with the exam board and invigilation errors during a recent administration of the Japanese National Unified College Examination (The test for Japanese final year high school students who wish to enter University, at a University in Tokyo). In 2012 550,000 applicants registered for the test and a total of 835 public and private institutions participated (Kyodo, 2012).

Building on the social concerns associated with testing practices, Gipps (1994) recognised the need for a paradigm shift (Kuhn, 1970) in order for stakeholders of examinations to consider assessment in new ways which are suitable for our modern day purposes and understandings. With this aim in mind, the following measures are suggested:

- A movement away from the norm-referenced limiting forces of psychometric models which emphasise ranking
- production of descriptions of performance which allow for idiosyncratic approaches to learning

- design of new and richer methods of presenting results using qualitative descriptors rather than merely figures
- generation of new ways of ensuring reliability and validity due to changing nature of assessments and their contexts
- consideration of ethical issues in the development and use of tests

Davies (1997) claims that language testers are, in general, genuinely interested in ensuring that the tests they work on follow ethical good practice. Furthermore, Bachman (2000) suggests that the solution to ethical misuse of language tests lies in codes of ethical good practice and this is perhaps a key message for Assessment Literacy. Such ethical guidance frameworks do already exist in the form of EALTA's Guidelines for Good Practice (Erickson & Figueras, 2010) and ILTA's Code of Ethics (ILTA, 2000).

In addition, (Kunnan, 2003) suggests a number of laudable principles for test fairness such as:

'Principle 2: *The Principle of Beneficence:* A test ought to bring about good in society
Sub principle 1: a test ought to promote good in society by providing test-score information and social impacts that are beneficial to society.'

The social and economic implications of testing are extended beyond the impact on local communities by Taylor (2009) who stresses the international importance of testing and assessment. Whilst countries with developed educational systems and established public examinations are, in many cases, now investigating further innovations in teaching and learning for the purposes of extending economic and social enhancement, other developing countries are at different stages and seek to follow in the footsteps of developed nations in order to harness

education and assessment methods and to benefit from the long term gain of their societies. As this is the case, the need for societies with developed assessment mechanisms to ensure that practices are sufficiently sophisticated and trustworthy is an even higher priority.

Nonetheless, despite warnings surrounding the risks associated with poor Assessment Literacy and the existence of models of good practice, there still appears to be a sense in the mind of critical language testers (McNamara & Roever, 2006; Shohamy, 2001) that the consciousness of the power of tests needs to be reawakened amongst language testers. Clearly, given that some of the most powerful contemporary tests are associated with EAP, this area of concern should have particular resonance for those seeking to develop EAP Assessment Literacy.

2.3.6.2 Preventing problems through pre-testing

The various stages of piloting and trailing that take place in development, before operational administrations of a test are deemed camera ready, can be described as pre-testing. Skill in this area can be seen to relate directly to the area of Assessment Literacy associated with the avoidance and prevention of test-related problems (Davies et al., 1999, p.150). As clarified by Bachman and Palmer (1996, p.234) the amount and kind of pre-testing collection will vary according to the purpose and scope of the test.

Fulcher and Davidson (2007, p. 81) compare the trialling and piloting of language tests to the manufacturing industry and refer to two stages of rapid prototyping which involves alpha testing and beta testing. Alpha testing refers to the in-house testing of pre-production parts to decide if design is suitable and to identify and remove any faults through expert judgement. In the context of language testing this could involve gathering the views of expert colleagues from the field of

applied linguistics. The second stage is 'beta testing' and this involves the external piloting of pre-production items with a representative group of tests takers. Nevertheless, the point is made that for large-scale tests that will be used to make inferences and decisions about large numbers of individuals, pre-testing is likely to be a more extensive and rigorous process. As many EAP testing contexts fall into the high-stakes category (Alexander et al., 2008, p.307) the need for an appropriate amount of pre-testing seems evident in order to avoid any problems which might influence the lives of test takers should their results on EAP tests not reflect their true abilities.

Bachman and Palmer (1996, p.234) suggest three phases of pre-testing as indicated in Table 11 below:

Table 11 Steps in pre-testing (Bachman and Palmer, 1996, 244)

Step 1	An informal stage where individuals (experts or representative students) and small groups provide qualitative feedback on a test regarding potential problems, specific tasks and instructions.
Step 2	The collection of more quantitative feedback on the test from larger groups of representative students or expert colleagues.
Step 3	A field trial conducted under operational conditions using the same administration procedures as when the test will be given for its intended measurement purposes. This field test stage however is only used to assess test usefulness and not to make formal inferences about test takers.

As explained by Fulcher (2010, p.159) the process of pre-testing can appear complex, however in high-stakes situations it is very important as the process is part of the evidence which can be used to support claims of validity for score meaning. With this in mind it is clear that understanding the processes involved in pre-testing can be considered an attribute of Assessment Literacy in EAP assessment.

2.3.7 Section conclusion

This section of the Literature Review has served to exemplify and contextualize key features and components of Assessment Literacy, with reference to pertinent cases drawn from the spectrum of literature that is relevant to the field of EAP assessment. In line with the research questions, the material selected has highlighted key research which has been undertaken that can be harnessed to

build and support Assessment Literacy for EAP practitioners. The next section of this review will explore other guidance resources and research which has developed and which arguably is relevant to the context of Assessment Literacy for EAP professionals.

2.4 Guidance resources for the development and enhancement of EAP Assessment Literacy

2.4.1 Section introduction

Given that the amount of literature, guidance and assessment training available to and designed specifically for classroom language teachers of EAP has been considered inadequate for some time (Brindley, 2001, p.127), it is perhaps unsurprising that over more recent years a series of more practical guidance tools have emerged with the precise aim of remedying this situation.

In this section, developments which have emerged over the recent history of language testing and which are pertinent to key aspects of EAP Assessment Literacy will be presented along with key resources available to EAP assessors from both the general language testing and EAP focused sectors. Once again, this addresses the thesis research questions by identifying research-led good practice associated with EAP Assessment Literacy from within these particular sources. It will also assist by highlighting and evaluating existing resources available to EAP teachers which could be used to enhance or sustain Assessment Literacy. Not all the research or resources discussed have emerged solely or directly from the context of EAP, however, those examples which will be referred to have been identified as they represent a useful and accessible potential resource for EAP practitioners who wish to enhance their Assessment Literacy.

2.4.2 EAP Assessment Literacy informed through reference to the history of language testing

In this section notable phases in language testing as a general field will be linked to associated developments in EAP testing in order to demonstrate key aspects of Assessment Literacy which are relevant to EAP teachers involved with assessment.

Table 12 (Masden, 1983 ; Spolsky, 1976, 1995) below presents certain key stages of language testing which account for particularly noteworthy developments within the academy since the 18th century.

Table 12 Key stages in the academy of language testing since the 1950s (Brown & Abeywickrama, 2010; Masden, 1983 ; Spolsky, 1976, 1995)

Approximate emergence	Stage of language testing
Since 18th Century and even earlier	Pre-scientific (traditional/intuitive) stage
From 1960s	Psychometric-structuralist (modern/scientific) stage
From 1970s	Psycholinguistic-sociolinguistic (post-modern) stage
From 1980s	The Communicative Era
The present day	Alternative Assessment Computer-Based Assessment Multiple Intelligences

Although the description of the different developmental stages within the history of language testing differs slightly between language testing experts (Moreno Espinosa, 2003)

Table 12 above serves to provide a summary of the stages which refer to historical developments and current debates.

The Pre-scientific Stage, also described as intuitive or traditional, is referred to in this manner as it involved testing by non-experts in the field, with arguably low levels of Assessment Literacy, using somewhat subjective assessment tools. Testing operating at this stage was not grounded on linguistic theory, and reliability but rather the priority was the creation of a test that felt fair

(Spolsky 1995, p.356). During this period when practitioners' instinct was the main resource, there tended to be a dependence on the assessment of grammatical knowledge (Masden, 1983, p.6). This era in language testing corresponds to the period in time for EAP when, in most cases, university provision for the needs of international students was also of an informal nature. According to Jordan (2002), any language support which was offered to international students usually took the form of ad hoc, part-time provision, as and when problems arose or developed during the course of studies. In terms of EAP Assessment Literacy, fostered through UK universities, it wasn't until 1971 that one of the first documented in-house EAP assessments was developed by Tim Johns at Birmingham University. Johns created a diagnostic EAP assessment for overseas postgraduate students.

As the label suggests, in the Scientific era, the psychometric-structuralist movement offered a more systematic and technical approach to language testing emerging through testing specialists with linguistic training and higher levels of Assessment Literacy. This stage harnessed the expertise of structural linguists who typically identified the linguistic focus for assessment and psychometric measurement specialists who were able to provide more objectives and reliable testing instruments (Spolsky 1995, p.353). This included sentence completion exercises and multiple choice items (Masden, 1983, p.6). This point in time aligns with the growing development in professionalism of EAP and its practitioners through the acquisition of Assessment Literacy. This situation coincided with the establishment of the professional body Special English Language Materials for Overseas University Student groups (SELMOUS). This organisation provided a forum for the sharing and enhancement of expertise of individuals such as Swales (1985), who were interested in English for Specific Purposes in the context of academic study. It should be noted however, that work of SELMOUS and latterly BALEAP has traditionally

specialised in the development of EAP teaching materials rather than assessment methods, as its original name suggests (Blue et al., 2000, p. 43).

In scrutinising the results of a BALEAP questionnaire into EAP testing practices Blue et al. (2000) showed that at the time a number of practitioners cited lack of staff time and expertise, insufficient time for test administration, absence of a requirement by outside agencies and a wish to avoid additional testing due to student and teacher familiarity with strengths and weaknesses in EAP. Blue et al. (2000) then go on to state that evidence now suggests that the situation is changing. This now certainly seems the case, given embarkations which show a higher level of commitment to and investment in EAP Assessment Literacy. An example of this includes the activity undertaken by the BALEAP Working Party on Testing (BALEAP, 2012a).

During the 1970s, a movement emerged which purported that discrete point testing such as that common during the scientific stage of language testing was insufficient for the multidimensional context of language learning and its assessment. Oller (1979, in Weir, 1990, p.56) contended that a more global integrative testing would represent a more accurate method of reflecting the need to combine language skills in a manner similar to their real usage. The applicability of the Multiple Choice Question (MCQ) format for EAP assessment has also since been challenged in the EAP context by Flowerdew and Peacock who claim that the format does not adequately test normal language processing skills. As support for this has grown, the psycholinguistic-sociolinguistic stage of language testing has emerged with the objective of realising a more unified approach to language testing, which would facilitate both functionally and socially contextualised testing of language use (Spolsky, 1995).

Gipps, (1994, p.16) also describes the need for a shift from the limitations of the psychometric model which emphasises ranking and statistically driven distributions towards a new model focusing on achievement. It is clear that critical and socially attuned language testing (Shohamy, 1998, 2001) can also be seen to have drawn support from this approach. In relation to EAP, this attitude towards language testing connects logically with the integrated skills method which aims to teach skills and create assessments whilst adopting a more unified approach to the underlying linguistic constructs. It is recognised that, in the real world, skill areas overlap and should therefore not be taught or assessed independently of one another (Jordan, 1997, p.141). This understanding, regarding EAP and the sophisticated use of language required in university contexts, acknowledges the need for students to be able to criticise a point of view and to evaluate different possibilities (Blue et al., p.8). In terms of Assessment Literacy in EAP, this is an important aspect of Construct Validity (See Messick, 1989). The critical aspect of the agenda associated with the psycholinguistic-sociolinguistic language testing also resonates particularly significantly with the field of EAP which has been linked with institutions associated with global high-stakes testing for a number of decades (Shohamy, 2001, p.123). This means that if assessments are unreliable or reveal only a very partial view of students' proficiency, then they may exclude students from entry to Higher Education or fail to predict very real difficulties which students may encounter (Blue et al., 2000, p.8).

One of the more recently observed stages in language testing is the communicative stage. Communicative language testing aims to collect information about the test-taker's ability outside the boundaries of isolated skills (Masden, 1983, p.7) as well as how language is used in real-world contexts. This stance also recognises that, irrespective of the authenticity of test tasks in relation to the test-taker's target usage domain, the test taker's performance will be constrained by the nature of the test-taking conditions. Weir (1990) can be seen as one of the pioneers of this approach. As

Weir (1990, p.6) explains, tests such as cloze tell us about a test-taker's linguistic competence but don't tell us anything directly about performance ability. One major criticism of Weir's focus on the communicative approach to language testing is that its preoccupation with content, authenticity and sampling does not address the deeper concerns about ability and competence which are now recognised as fundamental to construct validity (Fulcher, 1999, p.223). Further details connected to Weir's contribution to the TEEP test are discussed in section 2.6.2.

In order to bring a review of historical developments in language testing up to date, Brown and Abeywickrama (2010, p.17-21) have undertaken a review of contemporary issues and considerations which are currently in the process of influencing assessment practices in the classroom.

In his work on multiple intelligences, Gardner (1983, in Brown & Abeywickrama, 2010, p.17) identified eight areas of intelligence in the form of: linguistic, logical-mathematical, spatial, musical, kinaesthetic, naturalist, interpersonal and intrapersonal. An appreciation of these different intelligence types has subsequently had an impact on educators and their teaching and testing of languages. Consequently, many practitioners now strive to create assessments which take into account interpersonal, creative, communicative and interactive skills.

In EAP, this consideration of multiple intelligences sits comfortably within or alongside the concept of academic discourse communities (Alexander et al., 2008, p.6) which underpins much of the ethos behind a great deal of EAP-related pedagogy today. The concept of discourse communities acknowledges the fact that members of a particular academic school are likely to share understandings or a culture which differs markedly from those of other academic communities. An example may include students of particular university department who may

espouse quite differing research paradigms or ways of pursuing or communicating knowledge when compared to a sister department which is focused on another academic domain. ELT resource books such as Berman (1998) have assisted in supporting this view of teaching and learning languages which takes into account the differing learning paradigms pertaining to different individuals. Considering both these concepts allows us to remember that new members to an academic tribe have to go through a period of orientation before they can become experts. They have to learn tasks, behaviours and rituals. As they are human beings, they will also bring with them different affect or behavioural preferences and idiosyncrasies which will influence learning and how knowledge or skills can best be assessed (Popham, 2012a, p.5).

In the development of contemporary EAP assessments there currently also appears to be a movement towards alternative assessment methods which have been designed to supplement more traditional end-of course assessments. It seems that there is demand for a lesser focus on large-scale standardised testing and a growing interest in contextualised, student-centred performance- or task-based testing which allows for the evaluation of students while they are performing certain tasks. In the context of EAP, this is an ethos adopted by many contemporary pre-sessional Courses which focus on the development of international students' proficiency in EAP prior to the commencement of their main programme of study. These courses use a wider range of assessment types and are thought to give a more accurate indication of students' aptitude for university study than standardised proficiency tests (Alexander et al., 2008, p.308).

Given technological advances in recent years it is perhaps unsurprising that Computer Assisted Language Learning (CALL) and Computer Based Testing (CBT) has also developed as a field both in the classroom context through the availability of electronic teaching aids and through test-taker and institutional experiences with large-scale standardised tests which use internet or

computerised platforms. Some of the considerations associated with CALL which practitioners seeking to develop Assessment Literacy should be aware of are highlighted in Figure 9 below:

Figure 9 Advantages and disadvantages of computer-based testing

Advantages	Disadvantages
Easy to administer Can offer self-directed testing across 4 skills Practice for high-stakes tests Individualisation through Computer adaptive tests Large scale administration, marking and reporting of results with speed Much improved technology for speech recognition and essay marking	Lack of security, if unsupervised Some quizzes and tests available on the web may appear more formal or reliable than they are MCQs are prevalent and have potential for flaws Open ended responses are more complex to mark electronically Lack of human interactive element Validation issues exist as test-takers may not be able to approach tasks as a reflection of the real-world context

(Brown & Abeywickrama, 2010)

2.4.3 Key resources to enhance Assessment Literacy from the field of general language assessment

A seminal range of textbooks which spans a wide range of issues in the field of language testing, some of which also focus more specifically on contexts related to EAP, is the 'Studies in Language Testing' (SiLT) series (University of Cambridge, 2012b) which discusses a wide variety of key issues in language testing and assessment. Currently there are 34 titles available. These volumes deal with key areas in language testing, many of which fall into the following broad categories:

- IELTS and Cambridge English examination research
- Testing various language skills
- Test validation and analysis
- Test washback and high-stakes testing issues
- Investigating language policies frameworks and benchmarks
- Social and educational impact of language testing

Whilst this series undoubtedly offers a wealth of resource relevant to the EAP context, it has a varying level of accessibility and in many cases may require the reader to be more experienced in language testing.

A practical handbook for language testers is provided by Bachman and Palmer (1996) who describe the process of language testing through conceptual frameworks, the process of language test development and then offer exercises and tasks for local application. One particularly key contribution to Assessment Literacy could be considered to be the provision of a model for identifying test usefulness, first introduced by Bachman and Palmer (1996, p.38) and then revised by Douglas (2000, p.114) Table 13 below summarises Bachman and Palmer's original model:

Table 13 Characteristics of test usefulness Bachman and Palmer (1996, p.38)

Characteristic	Description
Reliability	Consistency of measurement and avoidance of variation in test scores due to factors other than the construct being measured.
Construct Validity	The meaningfulness and appropriateness of the interpretations that we make on the basis of test scores.
Authenticity	The degree of correspondence of the characteristics of a given language test task to the characteristics of a target language use task.
Interactivity	The extent to which the constructs we want to test are critically involved in accompanying the test task.
Impact	The various ways in which the test use affects society.
Practicality	The ways in which the test will be implemented in a given situation or whether in fact the test will be used at all.

Subsequently, Douglas's modification involved the conflation of authenticity and interactiveness into a single dichotomous category as it was felt that in the assessment of language for more specific purposes, such as EAP, there can be no authenticity without the presence of features of the target language use situation and the interplay with test takers' knowledge and the test task.

Davidson and Lynch (2002), Fulcher and Davidson (2007, 2009) and Fulcher (2006, 2010, 2012a, 2012b) have produced a series of commentaries and guidance tools which are particularly relevant and accessible for the development of EAP practitioner Assessment Literacy. Whether it be through the provision of procedural guidance tools, or the use of guided exercises, the range of

publications produced by these experts in particular has made important steps in breaking down barriers between theory and practice. Arguably, the most key achievement of these particular researchers, in terms of promoting Assessment Literacy, could be considered to be the demystification of construct validity so that it can be more readily understood and investigated by classroom practitioners.

In an alternative approach acknowledging the lack of time available to language teachers for formal assessment training opportunities, Malone (2011) presents a digest of online resources which includes a:

- directory of foreign language assessments (Center for Applied Linguistics, 2007)
- guide for assessment for foreign language educators (Center for Applied Linguistics, 2008)
- range of internet-based learning providing guidance with the development of proficiency-based language assessments (Center for Advanced Research on Language Acquisition (CARLA), 2012)
- website providing resources for general language testing (Fulcher, 2009c)

Perhaps the most dynamic and accessible resource listed above is the web-based material provided by Fulcher (2009b, 2009c) which offers a bank of information from the wider field of language testing as well as language testing in Higher Education contexts relevant to EAP. This site offers the user a range of media including downloadable articles and video clips from the media and language testing experts. Critical thinking is also encouraged through the introduction of scenarios for discussion and cases to consider.

2.4.4 Key resources to enhance Assessment Literacy in the field of EAP assessment

The series of articles presented by Blue et al. (2000) presents a range of important considerations in the context of assessing EAP which also, very practically, attempts to represent the needs of key stakeholders including academic departments and test takers themselves. This collection has a real world focus which is arguably very accessible to the EAP teacher at range of different levels of experience. The resource successfully integrates a combination of testing theory with the reality of the job in hand for EAP teachers working on Pre-sessional courses, looking for actionable solutions and realistic options.

Other useful and easily accessible tools for EAP assessment purposes have been developed by working parties operating under the auspices of the BALEAP, the global forum for EAP professionals. BALEAP's TEAP (Teachers of English for Academic Purposes) Working Party has compiled a framework of EAP teacher competencies (BALEAP, 2008) which is designed to act as a practice-led and expert-informed overview of the professional requirements of contemporary EAP practitioners. As might be expected, the document includes some focus on key EAP-assessment-related competencies.

Also very much connected the with work undertaken by BALEAP, Alexander et al. (2008) have provided a practical and user friendly EAP focused assessment guide as part of a classroom-focused teacher's guide. In the chapter devoted to assessment, the key attributes of both summative and formative assessment are investigated along with the implications of both formal and informal approaches. Accounts are given of some of the shared experiences of new EAP teachers who have struggled with the process of EAP assessment. The chapter aims to assist EAP teachers in bolstering their informal judgements with key principles relating to EAP assessment.

Interestingly, no attempt is made to encourage the use of statistics in the process of analysis and even a sceptical stance is taken with regard to the application, of such procedures for classroom practice. Instead, attention is devoted to the purpose of assessment and assisting EAP teachers in identifying the right test for the right situation. Further reference is then made to the characteristics of test usefulness as developed by Bachman and Palmer (1996, p.38) and extended by Douglas (2000, p.114). Advice, albeit brief, is provided on how to uphold the qualities of test usefulness as identified in this pre-existing model.

Another range of textbooks which offers a wealth of support to EAP teachers involved in assessment, is the *Cambridge Language Assessment* series (University of Cambridge, 2012a) which offers ten volumes focused on the assessment of different language skills such as writing, (Weigle, 2002) along with tools for statistical analysis (Bachman, 2004) and theory related to the use of computer based assessment (Chapelle & Douglas, 2006), a topic which is becoming more and more mainstream given the movement towards this medium by standardised tests. The contribution provided by Douglas (2000) on the topic of assessment in English for specific purposes also has particularly key relevance to EAP, given the familial linkage of EAP to ESP and the overlapping issues stemming from the need to define purpose and context, criterion and construct.

2.4.5 Section conclusion

This part of the Literature Review has identified a series of existing resources that are potentially accessible to EAP teachers and which may contribute to the process of informing good practice and enhancing or sustaining Assessment Literacy in the context of EAP. The next section of the

Literature Review will examine training opportunities which may also impact on EAP Assessment Literacy.

2.5 Training for teacher Assessment Literacy through general and EAP teacher education

2.5.1 Section introduction

According to Lukin et al. (2004), although an increasing amount of evidence supports the importance of assessment in the process of education, a number of researchers in the field of Assessment Literacy have documented widespread deficits in assessment skills evidenced by practicing teachers. As a result, in this section a review will be provided in order to consider training and educational opportunities which are available to EAP practitioners seeking to develop skills relevant to assessment. This section of the review can been seen as complying with the research questions from the perspective of investigating existing training interventions which can inform or be linked to the development of EAP Assessment Literacy.

2.5.2 Relevant assessment training material in general teacher education

According to Menter (2010), in a review of approaches to teacher education, there are four categories relating to the professionalism of teachers as described below in Table 14:

Table 14 Approaches to teacher education, adapted from Menter (2010)

The effective teacher	This approach positions the teacher as a facilitator of achievement and its measurement. This model frequently forms part of a national curriculum or assessment system.
The reflective teacher	The reflective practitioner embraces continuing professional development alongside teaching practice and characteristically involves a cycle of planning, delivering and evaluating teaching.
The enquiring teacher	Typically, enquiring teachers conduct classroom-based research, enhance their own practice and discuss their findings with other teaching professionals.
The transformative teacher	The transformative approach combines features of the reflective and enquiring models, in addition to knowledge transfer there is a social purpose at the core of this view of teaching.

Each of these paradigms focuses on different areas of emphasis within teacher education and attaches a varying level of importance to certain aspects of practice.

In many ways, the position of the '*effective teacher'* is arguably comparable to the function of EAP as seen by university management in many institutions. It also only encompasses a restricted form of Assessment Literacy. This stance places the EAP department in the role of converting otherwise unqualified international students into qualified students by developing language proficiency and finding mechanisms for the assessment and demonstration of this. In addition, the disestablishment of many university EAP departments and their replacement with external peri-academic 'partnership' colleges, suggests a mechanical/transformational approach rather than a scholarly objective with regard to positioning and management of such departments or units.

Whilst the development of knowledge and its assessment is key to the role of the EAP teacher and fundamental to Assessment Literacy, the practices of teachers which include the features described in the reflective, enquiring and transformative models provide a more holistic, dynamic and socially aware view of teaching which can be seen to be particularly relevant to EAP Assessment Literacy, given the need to take steps to uphold construct validity and keep a clear view of the needs of stakeholders who take and use tests.

Adams (2008, in Menter, 2010) positions the effective teacher, at the more narrowly skilled end of the teacher professionalism scale, with the progressive development of skill as the teacher moves along the reflective to transformative continuum and teachers become more proactive, enquiry-focused and critical.

This emphasis on the need to be proactive and reflective also resonates with the findings of Brown (2008) whose research in New Zealand suggests that increased levels of Assessment Literacy training does not necessarily result in teacher motivation to improve the quality of tests or assessments, especially when training focuses on assessment skills rather than on teachers' beliefs associated with assessment and the means of modifying them. From Brown's (ibid) research it seems that in-service experience of assessment has a more powerful effect on teachers' approach to assessment rather than any pre-service training undertaken.

Other materials for the training of teachers in assessment processes, from the field of general teacher education, include works such as Kyriacou (2007) who describes assessment as the series of techniques used to monitor student progress against specific learning outcomes. For Kyriacou the purpose of assessment can be summarised as follows:

- Providing teachers with feedback on student progress
- Giving students access to formative feedback
- Motivating students
- Recording progress
- Acting as a statement of attainment
- Determining students' readiness for future study

- Providing evidence of teacher and school effectiveness

Some of the drawbacks of assessment are also highlighted by Kyriacou (1991, p.111; 2007, p.105-6) who refers to Weston (1989) and five key negative aspects of assessment practice:

- An emphasis on norm-referencing, which causes students who experience difficulty to gain low marks and develop low self esteem
- A focus on summative assessment which provides grades but relatively little assistance in how to develop skills or enhance performance
- The prominence of cognitive and content-based study and the failure to recognise other types of accomplishment
- A power imbalance where students are subjected to assessment but have little control over the situation or understanding of criteria
- The aggregation of student achievement which can mask particular individual successes

Suggestions put forward by Weston (1989, in Kyriacou 1991, p.112) then argue that more attention should be devoted to formative assessment and that a wider variety of achievements should be considered with learning objectives more clearly specified. A more student-focused approach is also advocated which positions the student as a partner in assessment.

Kyriacou's (1991 and 2007) work, which has undergone numerous revisions, presents a useful basic handbook for educationalists across the curriculum; as a result, it can be seen as relevant to the EAP practitioner in an initial manner by providing assessment scenarios highlighting common pitfalls alongside practical advice for good practice.

In the 2007 iteration of Kyricaou's guide for trainee teachers, the training and development agency for schools' (TDA, 2007) Qualified Teacher Status (QTS) standards are referred to as these also provide guidance relevant to the Assessment Literacy of educators. These standards have been revised and reformulated in the form of the Department of Education's new Teachers' Standards document (2012) and refer to the following skills as part of making accurate and productive use of assessment:

- know and understand how to assess the relevant subject and curriculum areas, including statutory assessment requirements
- make use of formative and summative assessment to secure pupils' progress
- use relevant data to monitor progress, set targets, and plan subsequent lessons
- give pupils regular feedback, both orally and through accurate marking, and encourage pupils to respond to the feedback.

Although this set of guidelines is designed for trainee teachers training on courses such as Post Graduate Certificates in Education (PGCE), they still have relevance to the broader context of Assessment Literacy and resonate with the field of EAP.

2.5.3 Assessment training resources in the field of language teacher education

Despite the variety of materials available for the purposes of language teacher education, there is a very variable approach to the significance attached to the development of skills particularly associated with Assessment Literacy.

Some materials do already exist which make specific reference to assessement in contexts relevant to EAP (Alexander et al., 2008; Blue et al., 2000; Hyland, 2006; Jordan, 1997). However, the materials which are likely to be used by the newly-qualified English language teacher in the initial forming stages of their career have a varied approach to the emphasis and importance attached to assessment.

Two key texts which have commonly been used in initial teacher training over the last 20 years, and which have been updated and/or reprinted include titles by Ur (1999, 2012) and Doff (1988, 1995). Whilst both these titles arguably serve to provide a useful introductory function for trainee language teachers, neither places any considerable emphasis on the importance of assessment, although it is clear that the most recent iteration of Ur's work (2012) has enhanced this section, broadening the coverage to include assessment in the wider sense and making reference, in more than double the number of pages of the original edition, to functions of different types of assessment, grading, test design and administration. Ur's work does also invite some critical evaluation of the process of testing and its design associated with validity and reliability through the exercises which trainee teachers are invited to engage with.

In contrast, Doff's workbook simply presents the challenges of testing in terms of the basics of identifying constructs and choosing a test which aligns, seemingly quite superficially, with the construct in focus. Although test design and enhancement is approached in an iterative manner with ongoing improvement in mind, it is extremely limited in terms of impressing the importance of construct validity in a more in depth manner.

Harmer (2007, p.379-394), in his textbook often used by participants in Diploma courses in English Language Teaching to Adults (DELTA), extends the coverage of key considerations in

testing and assessment in a manner which is more useful and transferable for the trainee or in-service language teacher, although the significance is relatively underemphasised given its significance in most English language teaching contexts today. Nevertheless, some useful material is presented, including coverage of validity and reliability along with considerations such as item design and the concept of test washback. Different forms of assessment are investigated along with the context in which they are applicable. In addition, the characteristics of good tests are discussed and key considerations for the use of criteria and rating scales are explored.

The practice of testing is also only covered in a very brief and general manner in Scrivener's (2011, p.290-296) manual for trainee English teachers. The importance of validity and reliability is dealt with through a focus on objective and subjective marking, along with a set of criteria which determine the features of a good test. A set of scenarios is then presented which invite reflection on testing-related situations, both from the teacher/tester and test-taker perspectives. Whilst the material represents a useful springboard for further discussion it is far from comprehensive and fails to demonstrate the inherent complexities associated with language testing and assessment.

Stern (1983, p.436-440) includes assessment in his description of evaluation through a discussion associated with the essential components of curriculum. It is explained that intended learning outcomes and course design alone are insufficient, given the need to determine whether objectives have actually been met. The variety of possible functions of assessment are then presented by Stern (1992, p.353-357) in Table 15 below which highlights the many options and applications for assessment and in particular emphasises the distinction between formative and summative assessment.

Table 15 Types and stages of evaluation (Stern, 1992, p.355)

Stages of evaluation	Formative		Summative
1. Setting tasks/test items/test/battery of tests/examination	Informal – Day-to-day quizzes – Tasks – Slip tests – checks	Formal – Occasional class tests – Progress tests – 'mock' examinations	– Internal examinations – External examinations – Tests at mastery levels – Schemes of continuous assessment
2. Assessment in relation to criterion	– Student performance – Observation of process, inspection of product – Judgments: right-wrong, approval-disapproval, rating/marking/scoring, pass-fail, etc.		
3. Decision-making	– Re-teach? Explain? – Practice? New approach? – Remediation? Advances? (etc.)		– Promotion to next course? – Placement in appropriate stream or set? – Discontinue? – Professional promotion? – Change of status? – Entitlement? (etc.)

Wallace (1991, p.16) discusses how skills and good practice associated with teaching can best be learned as part of training courses for teachers of foreign languages and refers to the following models as described below in Table 16:

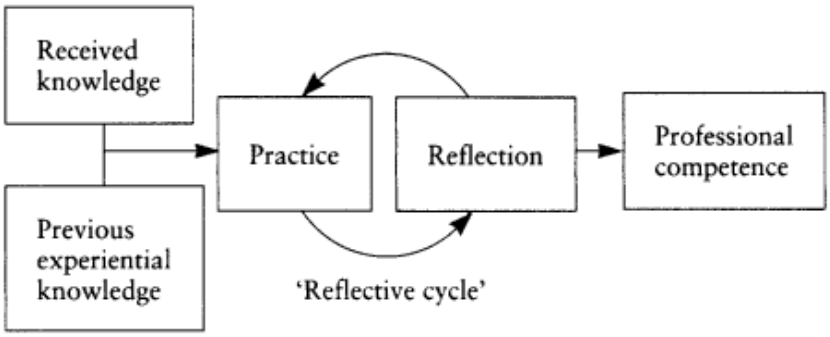
Table 16 Wallace's models of teacher education (1991, p.16)

The 'craft' model	Value is attributed to the experiential side of professional development in the same manner as an apprentice observing and emulating a master craftsperson. It is criticised as being closed to innovation and a mere act of mimicry.
The 'applied science' model	Takes into account the need to consider new knowledge and approaches but has been linked to a divide between teachings and research which risks jeopardizing the status of teachers.
The 'reflective' model	Is believed by Wallace to unify the positive elements of both the craft and applied science models, by giving equal consideration to both the experience of teaching and the research led findings which have emerged from researchers associated with the teaching profession.

Whilst Wallace (1991) supports this philosophy as a means for approaching and undertaking language teacher education as a whole, its significance is also particularly relevant to the development of Assessment Literacy for EAP. It can be argued that this is the case for the following reasons:

- It is important that trainee EAP teachers have practical experience of designing assessments and also are able to make use of relevant research from the field of language testing.
- If the hypothesis of this study is supported, then there will be some evidence to suggest that some EAP teachers who are involved in EAP testing and assessment have development needs which will require the acquisition of knowledge and the implementation of assessment good practice, linked to research.
- The 'Reflective Cycle' echoes advice provided by language testing experts (Fulcher and Davidson, 2007) regarding the iterative nature of language testing and assessment good practice.

Figure 10 Wallace's reflective model for teacher education (Wallace, 1991)



In a similar manner, Calderhead and Shorrock (1997) refer to the distinction between 'teacher education' and 'teacher training' by describing 'teacher education' in terms of the multifaceted education and development of teachers. In contrast, 'teacher training' is judged to be more similar to the craft model, as described above by Wallace (1991). The point is made that the 'teacher training' model is a more mechanistic method, emphasizing knowledge acquisition. On the other hand, the 'teacher education' model takes into account the need for teachers to think critically about their own practice and to justify strategies adopted in conjunction with students, given the different approaches possible and the need for flexibility in line with student needs and contextual

factors. This acknowledgement of the complex matrix which collectively comprises teacher education also supports the multidimensional challenges presented by EAP assessment as affirmed by both EAP and assessment experts (Hyland, 2006; Messick, 1989). Indeed, research from the broader field of teacher education (Calderhead & Shorrock, 1997) highlights the complex challenge facing teacher educators given the conflicting demands and limited time available for the purpose of teacher education programmes. Typically, a number of key dilemmas are presented which limit what can be achieved through such programmes. These dilemmas are now listed in Table 17 below and are contextualised with reference to their relevance to EAP Assessment Literacy:

Table 17 Key dilemmas in teacher education and their relevance to Assessment Literacy (Calderhead and Shorrock, 1997, p. 195-7)

Dilemma	Explanation and relevance to Assessment Literacy
Theory versus practice	This relates to the choices presented by the craft and reflective models of teaching as described by Wallace (1991) and the challenges of harnessing theory given the real world challenges presented by practice.
Content versus process	This choice is concerned with when to present certain essential skills in Assessment Literacy as they may be considered irrelevant if they are not linked to an opportunity for real-world application at the appropriate point for practitioners.
Gatekeeper versus facilitator	Tensions can be created as teachers of future teachers need to ensure skills are acquired but also have to act as mentors. This means that getting students to acknowledge weaknesses in Assessment Literacy may be difficult if students are aware of being assessed.
Personal development versus Professional development	Personal development is a key factor in teaching skill as teaching involves teacher's use of personality to manage interactions in the classroom. Personal development is therefore linked to the moral and social implications of testing and assessment; however personal development is difficult to define and to coordinate and is often sacrificed in favour of professional development.
Survival versus on-going development	Due to limitations on time available for courses, tuition may focus on strategies for the immediate survival of the newly qualified teacher or alternatively emphasis may be placed on the attainment of skills and understanding of theory which are not yet be relevant but which may arise or be drawn on in future situations. This means that certain aspects of Assessment Literacy may not be given adequate coverage.
Support versus challenge	Teachers need to gain challenging experiences through their education in addition to having access to the necessary support, the balance of how to manage these two aspects needs to be considered so that practitioners are simultaneously challenged and supported in their understanding of the complex range of capabilities which constitute Assessment Literacy.
Reproduction versus emulation	This refers to the challenge of preparing teachers to work in the institutions where they are required to work and function whilst also training them to be innovators. This aspect includes the skills of diplomacy and criticality so that flawed or imprecise assessment systems are not tolerated and yet people are able to work within constraints of existing teaching and learning contexts and cultures.

2.5.4 Training courses and CPD opportunities relevant to EAP Assessment Literacy

In recent years, the number of MA programmes and courses at other levels which have been introduced in UK universities for the purposes of providing training to teachers in EAP and Language Testing has increased. These courses have emerged due to practitioner concerns that practices by new language teaching professionals, including EAP teachers, who may typically hold a Diploma in English Language Teaching to Adults (DELTA) qualification, a sector standard for EAP teaching, are often deficient in terms of preparation for the particular challenges of the EAP classroom (Krzanowski, 2001; Roberts, 2001; Sharpling, 2002). This section will give consideration to learning opportunities both in the form of EAP-related study options and programmes which focus more directly on language testing. Table 18 below provides a brief overview of some of the current EAP-focused training courses, obtained from a review of course-related material on university websites. The aim of this exercise is to consider the content and extent of coverage, specifically focused on testing and assessment.

Table 18 EAP training courses and content focusing on testing and assessment

Institution	Programme Name	Testing/Assessment Modules
Sheffield Hallam University	PgCert Teaching English for Academic Purposes	Core area (1 of 6): Skills development and assessment
Warwick University	MA in English Language Teaching (with a specialism in English for Specific Purposes)	Elective module: (1 of 7) (15 credits) Assessment
Warwick University	MA in English Language Teaching (with a specialism in Testing and Assessment)	Core module (30 credits) Professional Practice (Testing and Assessment)
Leeds University	MA Teaching English for Academic and Professional Purposes (TEAPP)	No testing or assessment modules listed
Nottingham University	MA Teaching English for Academic Purposes (TEAP)	Elective module:(1 of 6) Assessment
SOAS	2- week intensive EAP Teacher Training (full-time)- An Introductory Teacher Training Programme	One session (Within 50-hour timetable) - Assessing learners' EAP and subject-specific needs
University of Essex	MA English for Specific Purposes (Subject to approval)	Core Module: Assignment Writing And Dissertation Preparation with options from the Essex language testing modules as above
University of Southampton	Teaching English for Academic Purposes- Short, intensive course for English Language Teachers with an initial qualification (e.g. CELTA or Cert TEFL)	No reference to assessment

Although the series of courses presented is not exhaustive, Table 18 gives some indication, that the focus on assessment in the EAP training courses listed is variable, the courses referred to appear to treat assessment quite differently in terms of whether it is included as a core or elective component and in some cases assessment is not referred to at all. Although the overview available from material displayed on the internet is limited, it is noted that there seems to be an inconsistent approach to what needs to be taught with regard to EAP assessment. As a result, this further highlights the varying level of importance attributed by HEIs and academic departments to this key area of EAP Assessment Literacy. With reference to training opportunities for EAP teachers, Sharpling (2002) also states that whilst the series of qualifications available to EAP teachers is broad, it is difficult to find parity across the range, in particular with regard to practical components.

According to Sharping (2002), the difficulty with courses used for EAP training purposes, is that EAP teachers are required to acquire knowledge which is not directly available via training courses and which takes time to develop usually through experience in the workplace. Such highly contextual knowledge may be better obtained through experience and through a process of continuous reflection. This is due to the fact that EAP involves complex needs analysis and an understanding of different registers and discourses within the academic field. It is also noted that assessment procedures within academic schools and departments are very fragmented.

Sharpling (2002) then refers to Brookes and Grundy (1990, p.36) who explain that departmental culture within universities places stakeholders in the roles of ‘insiders’ and ‘outsiders’ which involves power relations, in which a cultural mismatch can frequently occur, sometimes with severe consequences.

With this challenge in mind, Sharpling posits an argument, that whilst courses of this type may be useful in raising the profile of EAP as a profession, delivery of courses in EAP, such as those delivered at MA level, may not even be desirable as the course may not be fully achievable, for the reasons cited above.

Another crucial aspect of EAP training opportunity is presented by Allwright (1997) who highlights the importance of induction for newly appointed EAP teachers as often some staff embarking on a career in this field have limited experience particular to EAP rather than ELT. New staff may also be unfamiliar with the University and the idiosyncrasies of different departments and their academic demands. The key function of CPD and support for EAP teachers is also referred to along with the need for liaison and mentoring amongst EAP colleagues and interaction with other schools and departments within institutions.

Given the perceived need for skill development in EAP Assessment Literacy, another training option available to EAP teachers is to study language testing directly rather than to rely on language testing modules or components within EAP-related courses (Sharpling & Sky, 2010). Table 19 below offers an overview of some of the current language testing training opportunities at postgraduate level:

Table 19 Language testing training opportunities

Institution	Programme/Course Type Available	Course modules or components
Warwick University	MA in English Language Teaching (with a specialism in Testing and Assessment)	Core module (30 credits) Professional Practice (Testing and Assessment), including a site visit to University of Cambridge - ESOL Examinations
University of Essex	MA Language Testing and Programme Evaluation	Compulsory modules: Assignment Writing and Dissertation Preparation Communicative Language Testing Introduction To Quantitative Research Methods In Applied Linguistics Introductory Statistics For Applied Linguistics Using SPSS Language Programme Evaluation Language Testing
University of Lancaster	MA in Language Testing	Language Test Construction and Evaluation Background to Applied Linguistics for Language Testing Issues in Language Testing Research Methods Statistics for Language Testing
Roehampton University	Research MPhil/PhD (NB no MA listed)	Assessing spoken and written performance Assessment and learning Quantitative approaches in assessment Assessment and technology The social and educational impact of assessment Assessing the receptive skills

Whilst the courses above in Table 19 do seem to cover a broad cross-section of key skills in language testing, including procedures associated with test design and analysis, there is a varying emphasis on the social impact of language testing and indeed only one of the institutions listed refers to this specifically. Furthermore, whilst such programmes are clearly designed to appeal to potential students from a broad spectrum of language testing contexts, given the global impact of tests in EAP, it might have been expected that some of these courses of study would include modules on the assessment of English for more specific purposes.

In their two accounts cataloguing available courses in language testing and their instructional content, Bailey and Brown (1996) and Brown and Bailey (2008) describe and critique key training developments in the establishment of learning opportunities associated with the assessment, testing and measurement of language proficiency. It should however be noted that the focus of

these two studies is restricted to formal tests and test items and does not attempt to investigate any other form of assessment or other issues associated with assessment (Inbar-Lourie, 2008).

The results of these two studies demonstrate what is taught on some language testing courses and which skills and knowledge areas are considered important with particular reference to:

- Hands-on experience
- General topics,
- Item analysis topics
- Descriptive statistics
- Test consistency topics
- Test validity topics

The results of the most recent iteration of the questionnaire which was constructed demonstrate that some areas are new and dynamic whereas many similarities in response compared across both iterations suggest a stable level of knowledge rather than a dramatic shift in direction. Unfortunately, all in all, although the reader is invited to draw his or her inferences from the responses provided to each of the survey items, a synthesis which draws together the key findings and tracks the differences between the two surveys in 1996 and 2007 is notably absent and leaves the reader somewhat frustrated.

In addition to formal courses and programmes of study in the fields of EAP and language testing, a number of key opportunities and resources have emerged over the history of EAP and language testing which mirror the trends and developing level of professionalism in the academies of EAP and language assessment.

With regard to opportunities for the sharing of good practice amongst EAP professionals, professional bodies such as BALEAP (originally SELMOUS) (Jordan, 1997) and IATEFL emerged in the 60s and 70s. The Language Testing Research Colloquium (LTRC) and the Language Testing Forum then began in the late 70s (ILTA, 2012) and early 80s (LTF, 2012) respectively. The establishment of ALTE in the late 80s and 90s (ALTE, 2012) provided a practical mechanism for the setting of quality standards for language assessment . More recently, the Common European Framework of Reference (CEFR) (EAQUALS, 2012) was developed through the Council of Europe (2003). Resources and fora such as this have effectively created special interest groups and provided a vehicle for the enhancement of language testing and assessment procedures, including those in the context of EAP, outside the confines of individual institutions. More recently, good examples of this include the work of The International Language Testing Association (ILTA, 2000) which has produced a code of ethics for language testing. The European Association for Language Testing and Assessment (EALTA) (Erickson & Figueras, 2010) have also developed a good practice guide.

Another key resource for EAP teachers seeking to build Assessment Literacy has been the opportunity to train as an examiner for standardised testing systems; this has included tests such as IELTS and TOEFL which have introduced EAP teachers to key processes in standardised testing which they have then been able to transfer to their local contexts. Exposure to assessment methods in this way has provided EAP practitioners with a key awareness of the standards students attain on standardised EAP-style examinations, which are frequently used for gate-keeping and admissions purposes. It has also arguably afforded many professionals a critical perspective which is likely to have resulted in the transfer of good practices and the tools to enhance reliability and validity in their own testing contexts.

Calderhead and Shorrock (1997, p.197) describe the wide range of differing choices which commonly arise in the design and delivery of teacher education courses and reflect on the competing pressures which can ensue. The rationale for the variability in content and approach is attributed to differing value positions held within the professions of teaching and teacher education. As Calderhead and Shorrock (1997) suggest, varying views exist with regard to what constitutes the ideal role of teachers. Even within the same community; this may lead to divergence in beliefs about how teachers should be educated and supported. As a result, a cohesive conceptual framework which adequately accommodates and promotes the various theories and skilled practices of teaching seems to be absent or notably lacking for the purpose of teacher education and training.

2.5.5 Challenges in bringing the theory of assessment to EAP practitioners

2.5.5.1 Alienation of classroom practitioners

Brindley (2001, p.127) addresses a key aspect of the challenge in developing language teacher Assessment Literacy. Brindley (2001) maintains that a lack of prominence attributed to the process of assessment in teacher education courses, together with the density of language testing literature, effectively alienates the everyday classroom practitioner. In Brindley's view, testing literature often concentrates heavily on theoretical and statistical approaches to test validity and reliability. It is also noted that much of the literature is written from the perspective of research academics and measurement specialists, rather than teachers. Consequently, a situation arises whereby some teachers engaged in the process of language testing may feel that it is an excessively complex practice which is accessible only to experts (Taylor, 2009, p.21). Indeed it may sometimes be the case that assessment-related training is considered off-putting to some

potential users due to a perception that it is likely to be excessively technical or quantitative in nature (Stiggins, 1991). Popham (2009) also describes how beliefs associating assessment expertise with 'psychometric exotica' can be demotivating for some language teaching professionals. Brown and Bailey (2008) have also recently refreshed their investigation into the structure of existing training courses for teachers in language testing which again highlights certain key requirements for the improvement of training provided, so that it engages participants more effectively.

2.5.5.2 Privatisation of EAP departments

Given this recognition of the need for high quality training opportunities, it should be recognised that the training and professional development for EAP teachers, which forms a key aspect of Assessment Literacy, has arguably been jeopardised in recent times at many UK institutions due to the privatisation and commercialisation of EAP provision (Ansell, 2008). Over the last decade, an increasing number of British universities have opted to develop strategic partnerships with private companies which recruit international students and deliver EAP courses, resulting, in some cases, in the employment of less-well-qualified EAP teaching staff (Wright, 2012) and less favourable terms and conditions for language teaching professionals. In some cases, it is argued that such private providers bring an associated risk of reduced academic standards and flawed quality assurance procedures due to overambitious recruitment targets (Fulcher, 2007, 2009a; Hamp-Lyons, 2011; Lipsett, 2008). A detailed and politically impassioned account of the opposition to privately provided initiatives associated with teaching EAP is available from the University and College Union (2012), however it should be noted that the stance is not fully objective in nature. Hamp-Lyons (2011) describes, in a more balanced manner, the dangers of re-classifying EAP as 'professional' or 'support' whilst removing its connection with research at a

time when the fees from EAP activity are of crucial importance to the UK HE sector. Evidently, there are clear negative implications for the support and sustainability of Assessment Literacy for EAP practitioners if institutions proliferate which do not position academic professionalism as central to EAP good practice.

Despite recognised shortfalls in mechanisms for bringing Assessment Literacy to language testing and assessment practitioners, it should be noted, that there are some excellent ventures which have sought to rectify the situation. Recent examples which prove this include, Fulcher (2012a) who has investigated the general requirements of language teachers with regard to assessment needs through the use of a survey tool with innovative design features. This model has been particularly informative both in terms of results and methodology. Although it should be recognised that Fulcher's survey does not explicitly focus on the situation facing EAP teachers, whose roles and careers are arguably more closely bound to the challenges of high-stakes testing (Bruce, 2011) than colleagues in the wider and more general field of ELT, it does still provide a very useful exemplar of classroom focused research into Assessment Literacy. Another recent instance of good practice includes Walters (2011) who trialled Assessment Literacy tools with language teachers to gauge their bearing on the development of assessment-related skills.

2.5.6 Section conclusion

In this part of the Literature Review chapter, training opportunities have been discussed which can inform an understanding of how Assessment Literacy is fostered through more formal education systems. This has allowed the identification of educational experiences which EAP practitioners are likely to have encountered and other training mechanisms which may provide alternative models. The process has identified some sources of good practice. Challenges and potential shortfalls in available education and training mechanisms have also been recognized. This may be

useful during subsequent stages of the research, when considering whether the assessment practices of EAP teachers demonstrate Assessment Literacy.

2.6 University-led EAP Assessment Literacy through in-house testing

2.6.1 Section introduction

In addition to the learning resources available to EAP practitioners in the form of published materials and training opportunities, a number of UK Higher Education Institutions (HEIs) are also experienced in the development and usage of in-house EAP tests. Some of these institutions have also tried to compete in the larger scale market whilst others have focused on smaller-scale internal assessment requirements. This section will review two key in-house EAP tests in order to consider their contribution to the available resources for EAP Assessment Literacy. The impact of recent governmental interventions in in-house EAP testing will also be considered. In line with the research questions for this thesis, a review of how in-house EAP tests have been employed and developed over time also provides some insight into how the practices of EAP professionals have been influenced by research into language testing.

Table 20 below lists a number of the EAP tests, past and present, which are associated with UK HEIs:

Table 20 List of in-house EAP tests from UK universities

Acronym	Full test name	Institution
BrunELT	Brunel English Language Test	Brunel University
DALT	Durham Academic Language Test	Durham University
KITE	Kent International Test of English	University of Kent
OELPT	Online English Language Proficiency Test	Anglia Ruskin University
TEEP	Test of English for Educational Purposes	University of Reading
TELAS	Test for the English Language Assessment Service	University of Bedfordshire
UBELT	University of Bath English Language Test	University of Bath
UELT	University English Language Test	University of Leeds
USELT	University of Sussex English Language Test	University of Sussex
USEPT	University of Sheffield English Proficiency Test	University of Sheffield
WELT	Warwick English Language Test	Warwick University

Concerns relating to the quality of in-house testing systems are voiced by Coniam (2009) who stresses that for such assessments to correctly measure the abilities of test takers, training for test creators in the areas of test design and interpretation needs to be significantly enhanced.
Blue et al. (2000, p.8) support this view and warn that university in-house testing risks being unfocused and may vary in construct validity from institution to institution.

Brown (2008) also makes reference to the fact that many teachers have a limited understanding of the technical processes associated with the practice of principled assessment, this is substantiated by other experts in the field such as Stiggins (1995), Popham (2001) and Cizek (1995) who describe the challenge facing the classroom teacher of managing the complex matrix of skills and associated activities which constitute Assessment Literacy in the contemporary context.

O'Sullivan (2011, p.265-270) acknowledges the potential threats to test quality and claims relating to levels of student proficiency if teachers'Assessment Literacy is limited during the development of local tests. Certain key points are also made about the benefits of the local contexts in which many in-house tests are created. Rea-Dickins (2011) in O'Sullivan (2011) argues that in most situations local tests are more likely to allow practitioners to make valid assumptions about test-takers. It is also suggested that local test developers have opportunities to focus on specific domains and contexts which fit with local needs.

In the UK, the two tests about which arguably most is known in the public domain are TEEP, from the University of Reading and WELT from Warwick University. Developments surrounding these two tests will now be investigated further in order to highlight examples of EAP Assessment Literacy in practice which today's EAP assessment practitioners can potentially learn from.

2.6.2 Assessment Literacy through TEEP at University of Reading

In the 1980s Weir (1983, p.5) began work on the Associated Examining Board's Test in English for Academic Purposes (TEAP) the management and administration of which was subsequently adopted by University of Reading's Centre for Applied Language Studies (CALS) (now known as the International Study and Language Centre (ISLC). TEAP was subsequently renamed the Test of English for Educational Purposes (TEEP) (Weir, 1993). Weir along with Clapham (1996, p.4) and other developers of ELTS, a contemporary of TEEP, had a joint research interest in designing EAP tests which were based on authentic communicative situations required of international students (Weir, 1983, p.5). In particular, the direction of Weir's research was driven by desire to learn more about the language needs and communicative difficulties facing overseas students across discipline areas (Weir, 1983, p.547).

Weir's line of enquiry allied with that of Carroll (in Alderson & Hughes, 1981, p.67) who believed that content validity could be enhanced in the testing of EAP through investigating test takers and their communicative needs in the academic context. The main means of establishing content validity was through authentic sampling and representation of the disciplinary domain (Fulcher & Davidson, 2007, p.6).

Weir sought to explore the viability of harnessing his findings in order to construct a test specification (Waters, 1996, p.42-44) which could be used to create a test founded on the communicative reality of university study (Weir in Hughes, 1988, p.46).

In this respect Weir's contribution to Assessment Literacy in EAP is predominantly linked to criterion sampling and the principled use of a test specification for test design purposes.

The challenge for Weir, presented in identifying the communicative acts experienced by and required of international students is illustrated through reference to the framework which was used (in Hughes, 1988) for this purpose (See Table 21 below).

Table 21 Framework of categories for the description of communicative test events (Weir in Hughes, 1988, p.47)

Phase I General descriptive parameters of Communication	Phase II Dynamic Communicative Characteristics	Phase III Task Dimensions
Activities Setting Interaction Instrumentality Dialect Enabling Skills	Realistic context Relevant information gap Inter-subjectivity Scope for development of activity by participants Allowance for self-monitoring by participants Processing of appropriately sized input Normal time constraints operative	Size of text Grammatical complexity and range of cohesion devices required Functional range Referential range

According to Weir (in Hughes, 1988, p.73; 1983, p.547) one weakness of the TEEP specification was the need for more precise methods for dealing with task dimensions, that is to say that the extent to which different versions of the same test could be produced which differed only in terms of the subject context was questionable. In other words a realisation was developing that authentic sampling alone could not reliably lead to validity. Indeed, one major criticism of Weir's focus on the communicative situation is that its preoccupation with content, authenticity and sampling does not address the deeper questions about ability and competence which are now acknowledged as key to construct validity (Fulcher, 1999, p.223). Furthermore, the difficulty of accurately identifying what constitutes 'authentic situations' was also acknowledged.

Nevertheless, in contrast to Clapham's (1996) later rejection of the need for subject specific EAP/ESP tests, Weir contends that although the case for tests of English for Specific Purposes

remains unproven, attempts to describe what constitutes EAP across disciplines and the construction of valid and reliable measures for specifying students' proficiency remain worthwhile (Weir, 1983, p.549).

In addition, although authenticity is not usually cited as a critical quality of language testing, Bachman and Palmer (1996,p.23-24) argue for an important link between authenticity and construct validity by identifying that authenticity determines the level to which score inferences can be considered representative of the target language use domain.

Many of the original and underlying principles of Weir's work, undertaken more than 20 years ago, such as needs analysis and the linkage between reading and listening texts still apply to more recent versions of TEEP (University of Reading, 2004) which continues to be administered through the University of Reading's ISLC (Weir in Hughes, 1988, p.74). Weir's research is significant as it represents a design for a proficiency test based on an empirical understanding of needs. The central view that EAP tests should flow as naturally from needs analysis as the EAP course itself is however challenged by Fulcher (1999, p.221) who argues that concerns for authentic content should not obscure preoccupation with the main question of how valid inferences are drawn from test scores. Nevertheless, there is still an interest in learning more about the specific academic registers which university students have to deal with; this can be seen in studies conducted for the purposes of TOEFL (Biber et al., 2004).

Consequently, for practitioners interested in the development of Assessment Literacy, much can be learnt from TEEP and Weir's research both with respect to key contributions and limitations (Waters, 1996, p.43). In recent years, and with the wider scope of construct validity in mind, TEEP has been subject to a range of statistical analyses for the purpose of validating inferences

drawn from test scores. According to O'Sullivan (2007) in many ways TEEP could be considered a model for good practice which could be adopted by other Universities or colleagues seeking to test their own in-house EAP assessments.

2.6.3 Assessment Literacy through WELT at Warwick University

WELT was initially introduced at Warwick University during the 1980s for the purposes of assessment during in-house EAP pre-sessional courses. According to (Sharpling, 2010) the test was trialled and extensively and statistically validated. The writers and architects of WELT were English language teachers with international experience associated with language testing and assessment (Sharpling, ibid). In 1989 the test started to be used for direct admissions purposes at Warwick. In recent years, there was a rapid increase in the number of international students taking WELT. In 2005 there were in excess of 1000 test takers. For a period, in addition to Warwick test centre in the UK, WELT was administered in around 14 different countries in a series of 30 test centres appointed and overseen by the University of Warwick's Centre for Applied Linguistics, which managed WELT (Sharpling, 2007). Consequently, the acceptance of WELT, for direct entry purposes, spread across a wide range of UK HEIs.

Some of the key learning opportunities which EAP teachers looking to enhance Assessment Literacy can draw from activity related to WELT (Sharpling, 2008) may include:

- The test designers' consideration of 'affect'(Weir, 2005), namely maintaining candidate interest and motivation in test taking, this also treats washback and ethics as integral to validity, along with other factors such as administration procedures, the test environment; test-

taker characteristics, emotional state, and familiarity with the test (Van der Walt & Faans, 2008)

- Considerations relating to the use of Multiple Choice Questions (MCQs) to enhance test reliability, offset against attempts to avoid the reiteration of strong power structures through MCQs, given the fact that truth is absolute and predetermined by the test writer (Shohamy, 2001, p.24)
- Text selection for testing purposes and the use of an evidence based approach to validation (Weir, 2005)
- Practices associated with test trialling and item difficulty analysis

The above features align, to some extent,with Chapelle's views (1998; 1999) on construct validity, including the dual concerns of trait and behaviour. Both the trait and the method must be appropriate for the purpose of the test, thus referring to both the fundamental construct and the process of measurement (Weir, 2005, p.1).

Whilst this socio-cognitive model supported by Weir (2005), with an emphasis on 'content validity' (Fulcher, 2012b, p.356), has inspired recent iterations of WELT (Sharpling, 2010) and may be suitable to assist in test design for particular contexts, for the purposes of Assessment Literacy, the limitation with this view of validity is that the population of test takers that the test has been designed for may be so specific that it is not necessarily comparable to that of another testing context (Fulcher, 2012b, 356).

2.6.4 Government intervention with University EAP testing and limitations on Assessment Literacy

One of the key implications for universities seeking to use their own in-house testing systems for admissions purposes, since 2011, relates to a specific government policy concerning the types of test which are permissible and the need to demonstrate standards in relation to CEFR levels.

It is now no longer possible to admit international students to pre-degree programmes or pre-sessional language courses without a Secure English Language Test, which is listed as appropriate on the UKBA website. This single policy change has prevented the use of tests such as TEEP and WELT for Tier 4 study visa purposes and thus both tests have ceased to advertise this function at this level. WELT has been removed completely since 2010. This outcome demonstrates how the change of policy relating to language testing can impact a range of different stakeholders, including international students and scholarly activity at UK universities, it also changes the equilibrium in terms of knowledge and power in a manner discussed by Foucault. This modern-day example highlights the range of political considerations which have come to be associated with language testing and acts as a lesson to those seeking to develop Assessment Literacy of the human impact that changes in testing policy can have.

In addition to the impact on students' lives, there is a restriction which this situation places on the EAP profession in terms of mechanisms for developing Assessment Literacy through professional practice.

Language assessments have become one of the foundation requirements for citizenship in many countries and a key factor for immigration as well as long-stay study in a number of other countries. The principle of connecting language ability with citizenship is often based upon the

aim of developing national unity (Fulcher, 2012b, p.173), however, in the case of the tightening of restrictions for student visitor purposes in the UK, there is a markedly different rationale.

The UK government has pledged to reduce annual net migration to the UK by tens of thousands each year. As students are currently considered to count as migrants when they stay in the UK for more than 12 months, aims to reduce immigration now include the reduction of international student numbers (De Lotbinière, 2011; UKBA, December 2010). Consequently, regulations relating to language qualifications have tightened considerably since 2011 (Fulcher, 2012b, p.173)

Arguments against recent policy changes have been posited, such as the following viewpoint expressed by Patrick McGhee (in Swain,2012), chair of the Million+ University think-tank:

"It is tremendously powerful for students from all around the world to meet each other and understand each other's culture, and Britain has been really good at that… So it is very, very unfortunate that students have been turned away, that students are put off applying to the UK."

McGhee (in Swain,2012) favours the removal of international student numbers from immigration figures as it is argued that international students' aim is to return home after completing their studies. This view now also seems to be resonating with politicians. In a recent speech, Vince Cable, the current Business Secretary suggested (in Mason, 2012) that the current immigration policy may be "damaging to the perception of how we welcome talent from overseas".

According to Mason (2012) the UK government has been considering whether to exclude students' targets on reducing immigration. Indeed, by removing students from the figures the government could reduce immigration statistics by tens of thousands. Nevertheless, the concept of excluding students from these measurements has not yet been approved by Mark Harper, the

current immigration Minister, allegedly due to concerns that there may be a perception of altering the figures inappropriately.

If such a turnaround were possible, it may assist in allaying initial concerns, as expressed by UUK (2010), which have subsequently been realised, regarding the importance of having a robust immigration system which does not damage the success or indeed brand of UK university study (Meikle & Malik, 2012). This situation is further highlighted by the situation which London Metropolitan University recently faced (Grove, 2012) when it had had its highly trusted status revoked, partly due to reasons associated with student language proficiency.

2.6.5 Section conclusion

This section has presented an overview of developments associated with two well-known university-led in-house EAP tests. This has shown how institutions and their staff have developed skills in EAP Assessment Literacy through both assessment experience and related research. In connection with the research questions for this study, these case studies can be seen to represent both a valuable model and resource for other institutions and practitioners seeking to develop Assessment Literacy. As a result, it is hoped that recent governmental policy, which seems to favour the use of commercially available standardized EAP tests, will not inhibit local academic enquiry into EAP assessment and the associated acquisition of Assessment Literacy skills.

2.7 Chapter conclusion

Through this literature review, the key objectives which have motivated a number of the main supporters of Assessment Literacy have been discussed and explained. Features of Assessment Literacy have then been explored in further depth and pertinent examples have been provided from across the fields of Education, Applied Linguistics and EAP. An overview has also been provided of existing training opportunities which are relevant to the development of Assessment Literacy in EAP, along with additional resources which the researcher believes can act as guidance for EAP teachers and departments seeking to promote Assessment Literacy.

Having explored the domain of Assessment Literacy in relation to EAP, with reference to relevant literature, an understanding of the key aspects of this area of academic development can now be expressed in a manner which is relevant to the research questions in this study. As a result, the research tools, which will be discussed in Chapter 3, will be structured in order to investigate EAP teacher Assessment Literacy in terms of:

- training, skills, strengths and weaknesses in EAP testing and assessment
- approaches to EAP testing and assessment design
- views on/ experience of validity and reliability in EAP testing and assessment
- practices involving analysis and interpretation of the results and scores of EAP tests and assessments
- consideration of ethics in EAP testing and assessment
- influence of research and other resources for the purpose of EAP testing and assessment on EAP teacher assessment practices

Embedded within these areas of exploration will be mechanisms to identify features of Assessment Literacy amongst EAP teachers, for instance those listed in Table 2 on page 28. Such features have been identified in this chapter with reference to good practice models, professional guidance, training activities, development opportunities and expert opinion.

CHAPTER 3 RESEARCH DESIGN AND METHODOLOGY

3.1 Chapter introduction

Now that the theoretical framework which supports this study and its research questions has been explored in further depth through the Literature Review, in Chapter 2, the research design can be considered along with the appropriate methods for data collection.

This Chapter will commence by discussing the paradigmatic options and epistemological approaches which are at the researcher's disposal, before identifying a suitable paradigm and approach for this particular research project into EAP Assessment Literacy, given the requirements of the research questions. A rationale for the selection of data collection methods will then be described before further considerations for good practice in the use of selected research methods are discussed. At this point, the methodological design and related options will be presented along with an explanation of how the target population will be sampled and how the research tools have been piloted.

3.2 Paradigm options and epistemological approaches

3.2.1 Section introduction

A research paradigm is an overarching term which describes a researcher's epistemological, ontological and methodological principles (Denzin & Lincoln, 2005, p.22). In essence, the main reason why paradigms are referred to in research is that they allow reference to the means through which people examine or manipulate their environment (Stansfield, 2001). In this manner, a paradigm represents a series of logically connected conventions which guide and position the way research is undertaken (Bogdan & Biklen, 1992, p.33). In order to help identify what a paradigm encompasses for the purposes of this research project into EAP Assessment Literacy, reference

has been made to Table 22 below (Guba and Lincoln, 1989; 1994) which lists a series of three questions which clarify the matter:

Table 22 Questions characterizing the term 'paradigm' (Guba & Lincoln, 1989, p.82;1994;Mertens, 1998, p.6)

The ontological question:	What is the nature of reality?
The epistemological question:	What is the nature of knowledge and the relationship between the knower and the would-be-known?
The methodological question:	How can the knower go about obtaining the desired knowledge and understandings?

According to Gruber (1993), conceptualisation is key to the formation of every knowledge base as it allows the description of the elements, ideas and entities which jointly comprise a domain of interest. Consequently, an ontology can be considered to be a collective understanding of a set of beliefs or concepts which represent a specific field (Ding & Foo, 2002). The concept of ontology can be seen as particularly powerful when it is considered in terms of offering individuals or groups a joint framework of understanding. Ding and Foo (2002) extend this view when they describe the enabling properties of ontology which they believe offer an explicitly defined reference model thus improving the reliability and transferability of information. It should be noted however, that Gruber (1993) recognises that the way in which ontology is interpreted across different groups or communities can be problematic, given that interpretations of certain concepts may differ.

The theory of knowledge, or epistemology, is fundamental to philosophy (Cardinal et al., 2004), as it is concerned with the construct of knowledge itself and consequently has important implications for all academic domains and knowledge transfer situations. Epistemology posits the question, 'how do I know the world?' and 'What is the relationship between the inquirer and the known?'(Denzin & Lincoln, 2005, p.183).

Whether it is made explicit or not, all research undertaken is steered by one or more forms of theoretical orientation. Furthermore, any effective researcher will be cognisant of the theoretical foundation which guides their own particular methodology, as this will help to ensure a more systematic and impartial approach (Bogdan & Biklen, 1992, p.33).

3.2.2 An overview of major research paradigms

As described above, the philosophical direction of any research can have an important influence on the researcher's choices. Neverthless, researcher's choice of paradigm is not always straightforward, due to the varied options accessible for investigation and recognising, constructing or gathering knowledge.

An overview of the major paradigms which have supported, and in some cases restricted, contemporary research (Dörnyei, 2007) is now provided in Table 23 below. This will be used in order to present a summary of the breadth of options available and to consider which might represent the most suitable fit for this research into EAP Assessment Literacy.

Table 23 Overview of major research paradigms

Positivism/Post-Positivism (from 19th century)
Notes on ontology and epistemology
• Positivism disregarded knowledge claims which were not supported by scientific inquiry and promoted the view that society should be examined and researched through quantitative and empirical research methods similar to those used in the sciences (Law, 2007, p.40). • It was believed that positivism could explain public life through the inherent transparency of empiricism (Fuller, 2006, p.24). • Positivism claims that impartial accounts of the real world can be given and that reality can be revealed and proved (Denzin & Lincoln, 2005, p.27; Mertens, 1998, p.8). The researcher as the 'knower' and the truth as the 'known' are treated as separate factors (Tashakkori & Teddlie, 1998, p.7) Positivist research seeks to be objective and dispassionate (Guba & Lincoln, 1989, p.84). • In the 1950's academic cynicism led, to the appearance of Post-Positivism, which is based on a form of more critical realism and accepts that full objectivity is not possible, given that all research methods are restricted through inherent human limitations (Blaxter et al., 2006, p.59-60; Denzin & Lincoln, 2005, p.27).
Methodology
• Positivism and Postpositivism use predominantly quantitative or interventionist methods of research (Tashakkori & Teddlie, 2003, p.6). Frequently, approaches to research, similar to those used in the sciences, are applied to studies in the social world (Mertens, 1998, p.7). • Postpositivists sometimes make use of qualitative methods in the process of validating the results of quantitative studies (Blaxter et al., 2006, p.59-60).
Interpretivism/Constructivism (from 1970s)
Notes on ontology and epistemology
• Interpretivism or constructivism is founded on a basis of multiple socially constructed or voiced realities (Denzin & Lincoln, 2005, p.184; Mertens, 1998, p.8). • A crisis of representation and a genre diaspora was accompanied by increased interest in the multi vocal constructivist/interpretivist paradigm, along with qualitative methods. • Truth is considered to be the best informed and most refined construction of information on which there is a consensus (Guba & Lincoln, 1989, p.84). • Interpretivists or constructivists believe that people do not discover knowledge but rather construct it from various sources (Schwandt, 1994, p.125) and via interaction (Bogdan & Biklen, 1992, p.36). • In contrast to the positivists, this stance claims that it is not possible for the researcher to separate his or her values from enquiry. It is believed that there is an interactive link between researchers and their participants (Mertens, 1998, p.8).
Methodology
• Typically, the methodology is qualitative interactive and dialectical in nature. Interpretivists or constructivists often engage in expositive and hermeneutical research moving towards the analysis and interpretation of multiple sources and perspectives (Blaxter et al., 2006, p.59; Mertens, 1998, p.8).
Critical Theorism (from 1970s)
Notes on ontology and epistemology
• Critical theorists believe that multiple realities are shaped by social, political, and gender roles (Mertens, 1998, p.8). This approach challenges and criticises social and cultural subjugation and any perceived imbalance or misuse of power (Kincheloe & McLaren, 2005, p.304). • This philosophy emerged from frustration with other leading research paradigms which embody a potentially biased viewpoint (Mertens, 1998, p.15). This paradigm is informed by empathy with and support for marginal or oppositional social movements (Lather, 1992, p.88). • The critical and emancipatory philosophy challenges institutions which have a subordinating influence and aims to produce alternative approaches which uncover and challenge instances of inequality and injustice (McLaren & Giarelli, 1995, p.99). • Emancipatory and anti-discriminatory in nature (Creswell, 2003, p.138 and Denzin & Lincoln, 2005, p.18).For supporters of this paradigm research should seek to empower the powerless and lay claim, as Galileo suggested, to 'Science for the people' (Lather, 1992, p.88). Notably, participants in critical research are asked to consider the implications of their work beyond the confines of their current situation or project (McLaren & Giarelli, 1995, p.99).
Methodology
• Research methods for critical theorists are typically multi-faceted and diverse in nature. Although quantitative design can be utilised, qualitative methods tend to be prevalent (Mertens, 1998). Particular attention is paid to the avoidance of prejudiced or biased accounts. This Methodology frequently seeks to give a voice to individuals who are often marginalised in society.
Deconstructivism- Post-Modernism and Post-structuralism (from 1960s)
Notes on ontology and epistemology
• Postmodernists support the view that that no one instance of reality or truth should or can be fully defended as complete or irrefutable (Bryman, 2008, p.680). • All claims of truth are mistrusted as they are considered to support one or more particular objectives. Local, time-limited and contextually restricted accounts are preferred (Blaxter et al., 2006, p.61). • Poststructuralism involves language, subjectivity, social organisation and power (Richardson & Adams St.Pierre, 2005, p.961). Multiple truths and meanings are embraced (Urmson & Rée, 2005).

• In decontructivism, no individual stance can be acknowledged as correct without contest (Richardson & Adams St.Pierre, 2005, p.961). It is accepted that knowledge is affected by different viewpoints, outlooks or discourses and a person's conceptualisation cannot exist without some contamination from their own standpoint (Bogdan & Biklen, 1992, p.28). • The belief is held that historically accepted constructs and institutions are created through social rules and practices which regulate discourse (Law, 2007, p.342). As a result, knowledge is situationally limited. Deconstructivists prefer a situation where reality is jointly negotiated by the researcher and contributors. An element of reflexivity exists which accepts that a researcher's self and their research are intertwined (Richardson & Adams St.Pierre, 2005, p.962).
Methodology
• For deconstructivists, all research methods are equally distrusted (Richardson & Adams St.Pierre, 2005, p.961) and are open to enquiry and critique. Deep consideration of methodological approach is encouraged in order to encourage new methods of knowing.

3.2.3 Paradigm wars and paradigm supremacy

Given the fluctuation in popularity of different paradigms over time and context, the term 'paradigm war' has emerged (Gage, 1989, p.4-10) in association with assertions of paradigm supremacy and paradigm purity (Smith 1994, p.37-44). Indeed, in recognition of the challenges facing practitioners involved in language testing and assessment, McNamara (2003) has referred to paradigm wars which he suggests are 'tearing us apart again'.

Paradigm purists recognise the steering function of paradigms and, as identified by Greene (2008, p.70), emphasise that paradigms are constructed from interconnected and interlocking assumptions, which must be guarded as a complete set for the preservation of paradigmatic integrity. The mixing of axioms from different paradigms is actively discouraged (Guba & Lincoln, 2005, p.200-201). Furthermore, incompatibility purists such as Smith (1983, p.12) describe the incompatibility of conflicting positions which the different paradigms embody.

As a challenge to the concept of paradigm supremacy, Kuhn (1970, p.11) queries the degree to which any paradigm should be viewed as a fundamental or conclusive entity. This stance supports the view that a paradigm and its component elements should not be considered to be indisputable or compulsory. Kuhn's principal point is that a fully inclusive and watertight methodology, with the ability to settle scientific debate, does not and cannot exist (Von Dietze, 2001, p.4). In contrast, it is advocated (Kuhn, 1970, p.12) that the traditional and historical movement from one

paradigm to another, through revolution, forms part of the developmental pattern of science. Kuhn's research (in Fuller, 2006, p.19) claims that science adopts new paradigms through exemplary pieces of research which it then uses as blueprints for future inquiry. By adopting a particular paradigm, a common work pattern is agreed along with the means for supporting knowledge claims.

Kuhn's view positions most scientific research as a form of constrained adherence to existing models, as described by the blueprint for a particular paradigm. In Kuhn's view 'gestalt switches', or religious-like conversions cause scientists to search for a new paradigm during periods when unresolved academic or scientific problems or 'puzzles' accrue, as a result of a lack of confidence with the former paradigm (Von Dietze, 2001, p.5). Consequently, according to Kuhn (Von Dietze, 2001, p.5), theories and data of science rely upon whatever is the current dominant paradigm and with each paradigm shift, old knowledge is reinterpreted and seen in new manners.

Some challengers of Kuhn consider his interpretation of paradigm revolutions to be a imprudent form of unproductive liberalism (Woodhouse, 1996, p.5). Indeed, sometimes Kuhn's scientific revolutions are labelled as inhibiting the advancement of knowledge (Feyerabend, 1970, 197-198). This stance interprets Kuhn's key message as a caution against disposing of valuable approaches to research, while disposing of methods which may be considered outmoded or defunct in certain contexts. This point is highlighted in Kuhn's own conclusion that 'There are losses as well as gains in scientific revolutions, and scientists tend to be peculiarly blind to the former.'(Kuhn, 1970, p.167).

Many contemporary researchers share the view that paradigm commensurability is possible and even desirable, (Greene, 2008, p.53) and that perceived issues of incompatibility across different

paradigms are not as profound as some purists may claim. Datta (1994) supports the coexistence of paradigms and suggests a means of integration through a pragmatic approach. In addition, Gorard (2001, p.6) describes the false dualism of quantitative and qualitative approaches.

3.2.4 Pragmatism -a paradigm for this research into EAP Assessment Literacy

Pragmatism is generally regarded as an appropriate paradigm for mixed methods inquiry (Denscombe, 2003, p.116). It comprises a set of assumptions about ontology and epistemology which differentiates the underlying philosophy from positivism and interpretivism whilst including a number of the features advocated by critical theorists and decontructivists. A number of the axioms associated with pragmatism have been summarised below in Table 24.

Table 24: Features of the pragmatist research paradigm

Features of Pragmatism	
Ontology	Today's truth may not be truth tomorrow. Absolute truth is a hopeless cause.
Epistemology	Knowledge is provisional and based on practical outcomes. Knowledge is judged on usefulness and how it can be applied.
Methodology	There is no single best scientific method leading to indisputable knowledge. Perceived dualisms in research methods and different paradigms are seen as unhelpful. There is scepticism about the distinction between quantitative and qualitative methods.

As Duff (2002, p.22) acknowledges, in Applied Linguistics, a greater understanding and appreciation of different research methods can assist in developing the scope of research in the academic field. Furthermore, according to Dörnyei (2007, p.44) insight into complementary realities can be particularly advantageous in Applied Linguistics and it is clear to see that this is true of other instances where there is a need for critical thinking. For the purposes of this research project, where the views of EAP teachers need to be collected, an approach which enables the identification of complementary realities is seen as positive for the validity and reliability of interpretations.

Reichardt and Rallis (1994) even argue that the various philosophical paradigms available actually have more common factors than differences. In practice, it appears that researchers seem to combine paradigms and use a series of both quantitative and qualitative methods (Mertens, 2003, p.27).

Although early testing and assessment research, including work conducted by Cronbach and Meehl (1955), is often considered positivist in philosophy, due to its claim that only corroboration of hypotheses and the results of empirical research can provide sufficient evidence or meaning, the nature of some of this research also can be seen to align with the axioms of pragmatism, as recognition is shown for the fact that theories and methods are tenuous and temporal. Cronbach and Meehl argue that confidence in a theory is raised as more relevant evidence is gathered to support it. Contemporary testing research such as that conducted by Fulcher and Davidson (2007, p.10) tends to argue that the wider context of theories in language testing and language acquisition need to be kept in view as theory and observation cannot be separated. This means that a focus on quantitative methods alone will not serve the full breadth of testing and assessment related research.

When considering the research questions associated with this research into EAP Assessment Literacy, pragmatism also seems to present a practical solution which will enable the collection of data from EAP teacher views via different methods that have quantitative and qualitative features.

In response to the paradigmatic impasse which researchers into language testing and assessment may find themselves in when considering which philosophy to follow, McNamara (2003) asserts that that it is not the quantitative/qualitative choice which should cause the most academic

dilemma, as a number of important empirical studies into language assessment have utilised both types of method. In contrast, the key challenge for those involved in language testing and assessment research is the means by which the different components of Messick's classic validity framework can effectively be realised, See Table 4 on page 46 (Messick, 1989, p.20) in a way which ensures that the multidimensionality of the context and stakeholder needs are adequately taken into consideration. Evidence for the practicality of the pragmatist approach in the context of language assessment and testing are further exemplified with reference to Messick's Classic Validity Framework (Messick, 1989, p.20).

The upper two cells of the matrix stress the importance of collecting evidence in support of interpretations that are made on the basis of scores and test uses. Indeed a great deal of language testing research in this area involves the use of statistical methods. However the lower two cells highlight the importance of social and political values in language test design and usage. As a result, the implication is that the processes involved in maintaining and enhancing construct validity in language testing and assessment, a key set of activities associated with Assessment Literacy, cannot be completed solely through the window of a positivist paradigm. Rather, as Messick advocated, multidimensional validity research is required, and as a result, language testers need to consider the full spectrum of methods and approaches available. This leads back to the pragmatist paradigm.

An additional rationale for the adoption of a pragmatist paradigm for this particular research into Assessment Literacy can be linked to a description of approaches to construct validity by Chapelle (1998) who refers to different features of trait and behaviourist approaches in construct validity along with the underlying activities which might typically characterise their realisation.

Whilst trait and behaviourist approaches can be considered as two different models for construct validity, contemporary language testing researchers now acknowledge the importance of understanding how traits and contexts interrelate and this has led to a perspective known as interactionalism which incorporates both the language knowledge of trait theory and the contextual factors of behaviourism (Bachman, 1990, p.84).

This interactionalist philosophy provides further evidence of the multidimensionality of construct validity. In turn, as practices associated with maintaining and enhancing construct validity play a key role in the development of Assessment Literacy for EAP practitioners, the internationalist approach leads the researcher interested in Assessment Literacy to pragmatism as a suitable paradigm.

3.2.5 Section conclusion

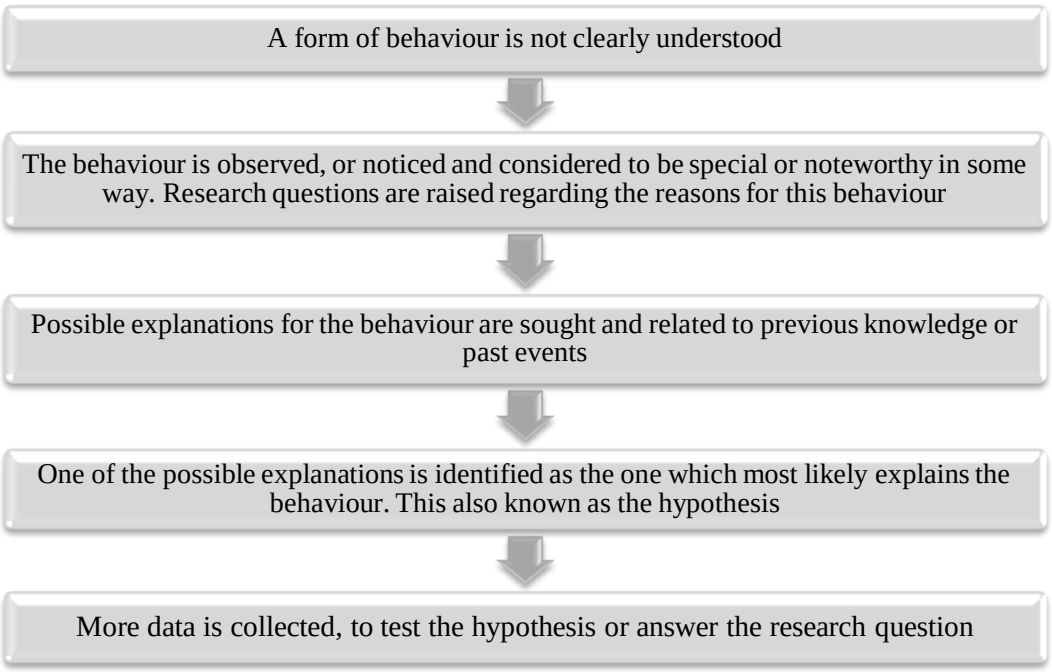
This section of the thesis has explored the paradigm options available to the researcher. Pragmatism has been identified as a suitable approach for this research into EAP Assessment Literacy, given the multi-dimensional overview which a mixed method approach can allow and the benefits to this project of investigating complementary realties. During this process the relative merits of common paradigms have been explored and concerns regarding paradigm supremacy have been challenged. Now that a paradigm for this study has been identified, it is possible to move on to consider the research methods which are most appropriate for the research.

3.3 Identifying research methods for Assessment Literacy and Applied Linguistics/EAP

3.3.1 Section introduction

Seliger and Shohamy (1989, p.7) clarify that all research activity is undertaken with the aim of finding answers to questions which affect our daily lives. Indeed this same logic can be applied to the research questions associated with this investigation of EAP Assessment Literacy. Some of the common triggers and features which lead to and characterise research are summarised in Figure 11 below:

Figure 11 Basic components of research in the real world (Adapted from Seliger & Shohamy 1989, p.7)



As depicted in Figure 11, it is the last component in the figure that describes the need for research methods. In the case of this research project, more data needs to be collected in order to test the hypothesis regarding EAP practitioner Assessment Literacy. The hypothesis and research questions are described in full on in Section 1.2 of this thesis, from page 21 onwards.

Research methods are the instruments for data collection that can be utilised for the purposes of, analysis, interpretation and explanation (Cohen et al., 2007, p.44). Consequently, this section of the thesis investigates the potential collection tools themselves as opposed to the product of research (Kaplan, 1973, p.23).

As it has already been acknowledged that pragmatism is the paradigm which has been adopted for this particular research context and project, it is therefore logical that both quantitative and qualitative research methods will be explored. Whilst Duff (in Chalhoub-Deville et al., 2006, p.66) advises that the schism between qualitative and quantitative methods is often exaggerated, it is also acknowledged that both quantitative and qualitative projects have their own potential strengths and weaknesses. As a result, in line with Dörnyei's (2007, p.18) approach to research methodology, the short review of qualitative and quantitative methods which follows will give an overview of the choice of research procedures (Seliger & Shohamy, 1989, p.155) which are available to researchers in the field of Applied Linguistics.

As inferred during the rationale presented for the selection of the pragmatic paradigm, although the dividing labels 'quantitative' and 'qualitative' are convenient, they often disguise a more complex reality. In fact, examples of research which are neatly separated entirely into quantitative and qualitative silos are relatively uncommon (Denscombe, 2007, p.119; Larsen-Freeman & Long, 1991; Reichardt & Rallis, 1994).

3.3.2 Identifying qualitative research methods

When investigating the potential qualitative research methods which could be employed to collect data for this particular research project, a series of possible qualitative tools were considered. The

table below shows why certain tools were rejected. Section 3.3.3 then introduces the research interview has been identified as a suitable option for the research required in this project:

Table 25 The rationale for rejecting certain qualitative research methods

Qualitative Research Method	Rationale for rejection
Focus groups	As the research being conducted is focused on collecting data from respondents across the EAP sector, a focus group would be impractical given time and geographical restrictions as well as the need not to focus on a group of practitioners from any single location or institution.
Observations	Given that the research questions require the collection of practitioner views rather than observation of behaviour, this method of data collection has not been deemed appropriate for this particular research project.
Documentary research and record review	This form of data collection has not been adopted in this particular research project, focusing on EAP Assessment Literacy, given the nature of the research questions which requires a more direct form of data collection from current and practising EAP teachers at a range of different institutions.
Case studies	Given the warnings associated with this research method associated with generalisation and the research questions which underpin this research project into EAP teacher Assessment Literacy which seek to identify trends which can be generalised to a wider community of EAP teachers, this means of data collection has not been selected.

3.3.3 Introductory rationale for selecting the research interview method

The research interview is one of the most prevalent and commonly identifiable forms of qualitative research (Mason, 2002, p.62). When information is needed in the context of teaching and learning, it is often expected that researchers will talk to the stakeholders involved (Drever, 2003, p.1). As a result the research interview can help to reveal the information and beliefs ‘inside a person’s head’ (Tuckman, 1972, p.196). As this qualitative research method is one of the data collection tools which has been selected for the purpose of this research project, this method will be evaluated in further detail in Section 3.4.

3.3.4 Identifying quantitative research methods

A number of possible quantitative tools were considered during the process of identifying suitable methods for the purpose of this research into EAP Assessment Literacy. The table below indicates the rationale for rejecting two quantitative methods which were considered. Section 3.3.5 then

introduces the questionnaire/survey method which has been selected as a suitable tool for usage in this research:

Table 26 The rationale for rejecting certain quantitative research methods

Quantitative Research Method	Rationale for rejection
Experiments	As this research method focuses on the relationship between different variables, this method does not lend itself to this research project which seeks to identify trends in EAP teacher assessment practice.
Tests	It has not been deemed appropriate to use a test for a range of reasons including practicality of administration and the fact that the aim of this research is not to evaluate performance in a manner which would reveal right or wrong responses (Dörnyei, 2007, p.103).

3.3.5 Introductory rationale for selecting the survey/questionnaire method

The survey or questionnaire is a self-report research tool which research participants are invited to complete (Tashakkori & Teddlie, 2003, p.303-304). Whilst the construct of questionnaires presents similar challenges to test design; the questionnaire gathers information from respondents for non-evaluative purposes (Tashakkori & Teddlie, ibid), there are no 'good' or 'bad' or correct or incorrect answers in a questionnaire. As the questionnaire has been identified as an appropriate tool for this particular research agenda, its usage will be critiqued and rationalised in Section 3.4.

3.3.6 Section conclusion

This section of the thesis has considered a range of qualitative and quantitative methods, which may be applicable for the purposes of research into Assessment Literacy, Applied Linguistics and EAP. Two model research tools, the questionnaire and the interview, have been identified as suitable for the exploration of the research questions which drive this particular enquiry. A brief introduction to the applicability of these methods has then been provided. A rationale has also been given to explain why other potential methods are less suitable to this research context. In the next section of this study it will now be possible to explore key considerations and good practice for the development of questionnaire and interview tools, designed to meet the requirements of this specific research agenda.

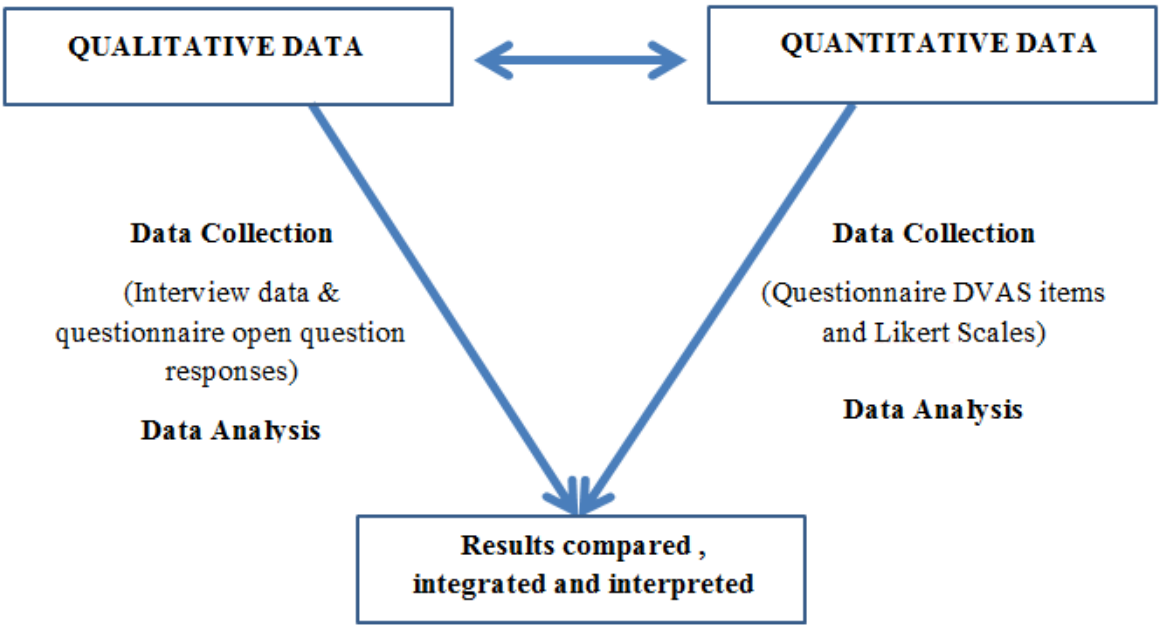
3.4 Designing and developing research tools for this research project into Assessment Literacy

3.4.1 Section introduction

The research in this thesis focuses on the collection of data associated with the Assessment Literacy of EAP teachers. The research questions which drive this investigation more specifically are listed and rationalised in more detail on page 21 in Table 1.

As clarified in previous sections of this thesis, the approach to research which has been adopted can be described as adhering to a pragmatist paradigm which will utilise a system of mixed methods. In this way, a form of methodological triangulation will also provide some additional support for the findings which are identified (Denscombe, 2007, p.138). In other words, the triangulation of quantitative and qualitative data collected via both questionnaire and interview methods will provide an additional level of corroboration for any findings (Dörnyei, 2007, p.165). A triangulation diagram which illustrates this objective is provided below in Figure 12:

Figure 12 Triangulation diagram



The particular empirical methods which have been selected for the purpose of this specific study into EAP teacher Assessment Literacy are summarised below in Table 27 above. In order to provide the requisite context, the research questions have also been included at the head of this table:

Table 27: Proposed data collection tools and their fit with the research questions

Research Questions:
1. To what extent do EAP teacher views on EAP testing and assessment practices reflect language testing research and practices which comprise Assessment Literacy? 2. How can EAP Assessment Literacy be sustained or enhanced?

Method A:	Questionnaire/ Survey
Qual/Quan/Mixed?	Mixed (Some closed some open questions)
Aim	To collect information about how EAP practitioners build EAP tests and assessments and to identify whether this information reflects key aspects of language testing research associated with Assessment Literacy
Analysis	Statistical analysis of responses which facilitate quantitative analysis and qualitative content analysis of emergent themes from open questions
Research Question	1 and 2

Method B:	Interviews
Qual/Quan/Mixed?	Predominantly Qualitative- (Semi-structured interviews which may include some closed questions for smaller scale statistical analysis)
Aim	To collect/ probe for further information from individual EAP practitioners on how they build EAP tests and assessments and to identify whether this information reflects key aspects of language testing research and demonstrates Assessment Literacy
Analysis	Content analysis of emergent themes from transcriptions of semi-structured interviews with open interview questions
Research Question	1 and 2

The design of the research tools utilised in this thesis is crucial in order for the study to be considered worthwhile, trustworthy and generalisable both in terms of the wider relevance of the findings which emerge and the relevance to others who may wish to conduct similar research themselves (Chalhoub-Deville et al., 2006 p.74). It is hoped that a mixed methods approach will enhance validity of inferences drawn from this study given the polarity of concerns such as those expressed by Adler and Adler (1994, p.377-92) on the one hand, who suggests that qualitative research is at risk of compromise given the intrusion of the researcher's role; and on the other hand, Blake (in Nesfield-Cookson, 1987 in Cohen and Manion, 2007, p.17) who argues that the quantitative approach, when alone, is reductionist and mechanistic in nature.

The following part of this thesis will build on the initial identification of the questionnaire and the interview as suitable research methods, as determined in Section 3.3. Further focus will be provided for the rationale for selecting the particular data collection tools identified above for the purposes of this research project into Assessment Literacy of EAP teachers. Strengths and weaknesses of each research method will be presented and key features of methodological design and piloting will be explained.

3.4.2 Considerations and good practice in conducting survey/questionnaire research

For the purpose of this research into EAP teacher Assessment Literacy, a questionnaire tool has been identified as an appropriate data collection method, as it provides a self-report research instrument which members of the target population can be asked to complete (Tashakkori & Teddlie, 2003, p.303-304). Efforts have been made to construct the questionnaire in a manner which complies with its formal research context and avoids the inclusion of faulty or leading questions.

Denscombe (2007, p.153) contends that a true research questionnaire should collect data which can be used for analysis and comprise a written list of questions which attempt to gather information by asking respondents directly.

In a similar manner to other research tools, there is a range of structural choices in the construction of the questionnaire, which may result in the creation of a tool which could be wholly qualitative, with open-ended questions or at the other end of the scale, include predominantly closed questions (Dörnyei, 2003; Seliger & Shohamy, 1989, p.173). For the purposes of more quantitatively focused questions, questionnaires often make use of rating scales which involve Likert scales and semantic differentials (Gorard, 2001, p.80). The questionnaire is also often used in conjunction

with other methods such as interviews in intermethod mixing (Drever, 2003, p.3) and therefore is commensurate with a mixed method approach, common in the pragmatist paradigm.

In the view of some quantitative researchers, such as Gorard (2002, p.190), it is sometimes difficult to defend the use of a questionnaire in research as the tool tends to be more efficient for the collection of simple facts rather than attitudes or opinions. On the other hand, qualitative researchers such as Mason (In Gillham, 2000b, p.1) claim that the added flexibility of open questions can be used to enhance the validity interpretations drawn from survey results.

Unfortunately, Dörnyei (2002, p.9) argues that no method has been more abused, partly due to the misconception that questionnaires can act as a rapid remedy to restrictions in time and other resources which researchers may be forced to contend with (Nunan, 1992, p.143). As Dörnyei (2007, p.102) clarifies, the design of a reliable questionnaire from which valid inferences can be drawn, is a highly complex procedure.

Despite the concerns raised by practised researchers, regarding the possible pitfalls of using a questionnaire, having evaluated the research context, the research questions and the type of data which needs to be collected, it is still believed that the questionnaire represents one of the most suitable research methods for the purpose of this research project.

Figure 13 below, has been devised in order to demonstrate how the inherent weaknesses of this data collection mechanism have been mitigated whilst strengths have been harnessed for the purposes of this particular investigation.

Figure 13 Strengths and weaknesses of questionnaire usage for this research

Strengths and means of harnessing	Weaknesses and means of mitigation
• **Good for measuring attitudes and collecting content data** This method has been efficient for collecting attitudes of a large group of EAP teachers but attention has been required to probe adequately in line with the research questions and to use appropriate question items. • **Fair reliability** Efforts have been made to choose and phrase questions carefully so as to collect rich and valuable data. Other research methods need to be used to support or refute inferences which are drawn from the data. • **Can be used efficiently with large groups of respondents** The researcher has benefitted from the opportunity to collect data from a wider group of respondents. An electronic survey has been used in order to reach as wide an audience as possible. • **Closed answer responses can be easy to analyse in volume** Some closed items will allow for statistical analysis, whilst other qualitative responses will require the use of analysis grids to highlight trends and patterns. • **Flexible use in conjunction with other methods.** The use of interviews will be complemented by collecting more qualitative data through interviews.	• **Inflexible, probing is not possible** Areas deemed to warrant further probing or further flexibility will be incorporated into interview question design. • **Often low response rate** Methods of increasing response rate have been considered, within the realms of objectivity and ethicality. • **Can result in selective non response to some items** The benefit of using an electronic questionnaire is that responses can be set as required in order to progress through the questionnaire. • **Open responses difficult and time-consuming to analyse** The researcher has developed a mechanism for content analysis this through the pilot research project undertaken prior to the thesis stage of this project. • **Questionnaire needs piloting and validating** The researcher has had access to a large group of local EAP teachers, so finding a representative pilot group who have subsequently been excluded from the main research has not presented difficulty.

(Inspired by Dörnyei, 2007; Drever, 2003; Gillham, 2000b; Gorard, 2001; Tashakkori & Teddlie, 2003)

3.4.3 Questionnaire design

Typically, questionnaires are complex structures used for data collection focusing on items which request quite specific information. For these reasons, the questionnaire is particularly well suited for quantitative data collection and statistical analysis (Dörnyei, 2003; 2007, p.104). In this manner it can be seen that the questionnaire is useful for the quantitative aspect of this research agenda, although a number of questions have been included which also allow for flexible open responses.

It is generally accepted that questionnaires seek to measure factual, behavioural and attitudinal information (Bethlehem, 2009, p.44; Denscombe, 2003, p.155; Dörnyei, 2003, p.5; 2007, p.102).

As a result, there are a number of considerations associated with the type of questions that need to be asked which will affect the design and structure of the questionnaire that is produced. One of the most common means of distinguishing items or questions in questionnaires is to categorise them according to closed or open variants. Question options which fall within the closed category are summarised as follows:

Table 28 Closed item types for consideration in questionnaires

Multi item scales & Likert scales	These scales refer to a cluster of several differently worded items that focus on the same target construct. The item scores for the similar items are summed resulting in a total scale score. The most well-known form of scale in this context is the Likert scale which includes a series of items with statements and responses indicating the extent of agreement.
Semantic differential scales	This method avoids the need to write response statements and respondents are asked to mark a response by placing a mark on a continuum between two adjectives.
Numerical rating scales	This involves awarding marks out of maximum in order to describe a feature, for example on a scale of one to ten how much do you enjoy EAP tests?
True or False items	These items involve a yes/no or true/false decision. Unfortunately, such forced decision items can distort data given the over simplification of the domain of choice available.
Multiple choice items	These may typically be used in questionnaires for personal information such as age, education or type of institutional affiliation.
Rank order items	These questions involve assigning a number to a list of items according to preference or perceived priority.

(Dörnyei, 2003, 2007)

Table 29 Open question types for consideration in questionnaires

Specific questions	These questions request particular information such as facts about past experiences or activities
Clarification questions	These questions ask for further details after a response to an earlier question or allow respondents to specify other contextual details
Sentence completion	Respondents are asked to complete a sentence in order to reveal how they feel about a particular matter or issue
Short answer questions	These solicit succinct answers to a question and are often longer than a sentence but shorter than a paragraph

(Dörnyei, 2003, 2007)

All of the question types, as listed above, feature in the final questionnaire tool to a greater or lesser extent, given the need to collect different types of information at different points of the survey. It should be noted however that the use of multi-item Likert scale questions is the most common question format used in the survey. The main rationale for the prevalent use of an inter-related scale of items is the recognition that when measuring a construct such as Assessment

Literacy and its sub-constructs there are problems if there is a reliance on isolated responses. This can be summarised as follows:

- A single indicator risks incorrectly classifying respondents
- One item may represent only a portion of the concept or construct being examined
- Finer distinctions can be made using a wider range of items which reflect various aspects of a construct or concept

The manner in which other types of question are employed, within the series in this questionnaire, can be viewed through examining Figure 15 on page 155.

3.4.4 Questionnaire medium - considerations for the use of an online survey

For the purposes of the questionnaire created for this research agenda, it was decided to use an online survey tool in order to be able to access a sufficiently wide number of respondents. Given the international implications associated with EAP and its assessment, it was also seen as a suitable means of capturing the data of practitioners outside the UK Higher Education environment.

Internet-based survey tools usually function by asking potential respondents to visit a website where the questionnaire is located, and which can then also be completed online. Whilst there are key advantages with using online media for the purpose of data collection, there are also methodological problems which need to be addressed in order for inferences to be made with respect to the target population (Bethlehem, 2009, p.277). Figure 14 below summarises some of the strengths and weaknesses particular to online survey tools and demonstrates how the researcher has sought to benefit from the advantages and minimise potential risks:

Figure 14 Strengths and weaknesses of online survey tools (Bethlehem, 2009; Bryman, 2008, p.653)

Strengths and means of harnessing	Weaknesses and means of mitigation
• **Low cost- in comparison to postal questionnaires although there may be some software-related charges**	**Low response rate- typically online surveys have a lower response rate than postal questionnaires.**
The license to use the selected survey platform is available through the researchers' employer and costs have been minimised through dissemination of the questionnaire via EAP-related electronic fora.	Efforts were taken to identify professional fora for EAP teachers which would allow dissemination of the questionnaire to the target population. An incentive mechanism involving Amazon™ vouchers was also used in order to encourage participation.
Faster response – online questionnaires tend to be returned more rapidly than other modes	**Restricted to online populations- only people with access to the internet can use this type of questionnaire**
In addition to dissemination of the hyperlink giving access to the questionnaire via email and online discussion boards, reminders have been sent by email to alert potential respondents and to improve the response rate.	Given the nature of the target population this factor was not considered influential as it was deemed that the majority of EAP would have access to the internet.
Attractive formatting for presentation purposes	**Requires motivation- if there is a cost implication in using the internet an incentive may be required.**
Visual impact and presentation has been enhanced through the use of colours and imagery.	Once again, given the nature of the target population it was considered that most EAP professionals would have access to the internet through their employment, in addition to any access for which they may commonly pay for private usage. Nevertheless, an incentive was still offered in the form of Amazon™ vouchers.
Mixed administration- can be combined with postal questionnaires, if the option is required.	**Confidentiality and anonymity issues- this may be a problem if respondents have to attach or embed their responses to an email which shows their identity but is less of a problem for purely web-based questionnaires.**
Whilst this option was considered, the large response rate to the online questionnaire did not result in the use of a postal version. Alternatively, attention was devoted to sending email reminders and using a range of different electronic professional fora for dissemination of the survey.	A purely web-based questionnaire was used in order to avoid the problems listed above. Whilst anonymity was guaranteed, disclosure of a contact mail address was requested at the end of the questionnaire in order to identify volunteers for interview. A significant number of respondents provided this information voluntarily and due care was taken during analysis not to allow this information to bias interpretation of results.
Unrestricted compass - no geographical constraints or problems returning data e.g. when responses are sent by post from overseas.	**Multiple replies- some people may seek to complete the questionnaire more than once and thus contaminate the data set**
To take advantage of this situation and the international nature of EAP, the survey was distributed via a number of international fora used by teachers of EAP.	Given the relatively time consuming nature of the questionnaire it was thought unlikely that participants would seek to complete the questionnaire numerous times, however, as a precautionary measure, a feature of the web platform was employed to prevent the completion of the survey from the same computer more than once.
Fewer unanswered questions- some evidence suggests fewer items are left incomplete	
In order to require answers to certain specific answers the questionnaire was programmed to prohibit progression if questions were left blank. This resulted in a large number of fully completed questionnaires.	
Better response to open questions- open responses seem more likely to be completed in online formats	
With this phenomenon in mind a number of open questions were included within key sections of the questionnaire in order to allow respondents to provide unrestricted open responses and to comment on any features of the questionnaire which they felt unable to do elsewhere in the questionnaire.	
(Bethlehem, 2009; Bryman, 2008, p.653)	

For this research project a number of online research tools were considered, these included, SurveyMonkey; Qualtrics, Lime Survey and Bristol Online Surveys (BOS).

Whilst SurveyMonkey has a particularly intuitive 'What you see is what you get' (WYSIWYG) user interface from the design and administration perspective and includes the useful feature of randomisation, there are a significant number of limitations for users of the free or lower cost versions. Lime Survey is a free and open source software option, which has a range of useful

features for design and analysis, importantly it requires installation on a webserver and familiarity with the associated technology, furthermore, the design and administration interface is far from intuitive. In addition, whilst Qualtrics appears to have the scope to be an exceptionally powerful and flexible tool, there is less scope for experimental trial before usage and payment of the required licence fee. Finally, it was decided that the BOS survey tool (University of Bristol, 2012) would be the most suitable option for this project.

Whilst the interface is not as intuitive or flexible from the design and administration perspectives, with certain convoluted processes for the design of certain questions and limitations on the range of options for visual presentation, its creation from within the academic sphere at Bristol University has led to embedding within the platform a number of key analytical tools which make BOS particularly useful at the analysis stage. It was also decided that, provided the researcher could invest time in gaining familiarity with the interface, the limitations as described could be overcome and the prospective respondents would not be adversely affected. One other key factor in the researcher's choice to use this particular online platform was the accessibility and affordability of the licence for BOS usage, available through the researcher's employer.

The features embedded within the BOS software are listed in further detail below:

When creating a questionnaire using the BOS software, there are several types of questions that can be created:

- Selection List
- Multiple Choice
- Multiple Answer
- Single Line (text box)
- Multiple Line (text box)
- Date
- Grid

The first item type, ‘Selection List’ provides respondents with a drop-down closed series of responses, only one of which can be selected. This type of item is useful in situations where the researcher wishes respondents to make a choice from a known list of options. Multiple Choice questions (MCQs) allow the creator of the questionnaire to ask respondents to select from mutually exclusive responses which require the selection of just one option. Questions in the ‘Multiple Answers’ item type offer respondents the opportunity to select as many responses from a list which apply to them or their circumstances. Text box question options are generally used for free-text purposes and allow longer or shorter open responses according to the size of the box. The date item is self-explanatory and can be used to ask respondents to identify a relevant date. One particularly key function is the final ‘Grid’ option which allows the designer of a questionnaire to combine all other question types through a table of rows and columns. Grid format questions are useful for the purpose of multi-item scales such as Likert scales or more generally, DiscreteVisual Analogue Scales (DVAS).

In order to ensure, as far as possible, that the questions which were designed for the research were fit for purpose the following steps were implemented (Bryman, 2008, p.241-244):

- The research questions, as listed on page 21, were kept in view in order to ensure that the questionnaire items accurately reflected the research objectives
- Questions were written with the solicitation of specific types of information in mind and questions were revised in order to attempt to ensure that they were not too general or obtuse in nature.
- Where possible ambiguous and technical terminology was avoided and where applicable complex concepts were explained

- Overly lengthy questions were avoided so that respondents are able to follow the key points without losing track or motivation
- Individual questions with double-barrelled elements were avoided as these are complex, if not impossible to answer in situations where only one response is made possible.
- Where possible, all questions which were initially formatted to include negative structures such as 'not' were rephrased in a positive form. This was undertaken as in some cases respondents can misread questions which are negatively phrased and the word 'not' this risking a misinterpretation.
- A considerable amount of attention was paid to the nature of each question and the scale or answer format which was provided, in order to ensure the request symmetry and balance of response options available. This was particularly important for the purpose of the items which collectively formed a Likert scale.
- Although a guarantee was given to respondents of their anonymity, certain biographical questions were asked at the start of the questionnaire in order to determine the level and extent of experience pertaining to each respondent. This was designed to assist in the filtration process during the data collection stage so that only those respondents who are deemed to have the requisite knowledge and experience were included in the analysis process.
- Question choice and phraseology was also reviewed by another colleague before being operationalized. See Section 3.4.6.

The full questionnaire, which was administered to respondents is available in the appendix of this document in section 8.1 on page 420. Figure 15 below provides an overview of the questionnaire structure and objectives of each element and question item. Further linkage of the questionnaire to the research questions is demonstrated through Table 30, which shows how areas of EAP teacher

Assessment Literacy, reflected in practice and research and summarised in the conclusion to the Literature Review in Section 2.7, are connected to items in the questionnaire.

Figure 15 Questionnaire structure and item features

(Page 1)

Welcome Page

- Welcome message
- Topic background overview
- Definition of assessment
- Details of incentive
- Context of study at the University of Leicester
- Instructions for survey page navigation
- Survey duration indication
- Overview of survey organisation
- Reminder of opportunity to opt in to survey participation
- Thank you message

Save & Continue to Next Page

(Page 2)

Data Protection & Ethics

- Confirmation of adherence to data protection protocols
- **(AQ1) FACTUAL** MCQ single answer **(yes/no) -** Participant agreement request for the use of data
- Hyperlink to download *Research Information Sheet* providing full details of the research project
- **(AQ2) FACTUAL** MCQ single answer **(yes/no) -** Participant confirmation of access to Research Information Sheet
- Invitation to email researcher with additional queries
- **(AQ3) FACTUAL** MCQ single answer **(yes/no) -** Respondent's acknowledgement of voluntary participation and facility to withdraw
- **(AQ4) FACTUAL** MCQ single answer **(yes/no) -** Agreement request for respondent participation in the survey
- **(AQ5) FACTUAL** MCQ single answer **(yes/no) -** Permission request for the researcher to use the data in anonymised form

Save & Continue to Next Page

(Page 3)

Personal Details

- **(AQ6) FACTUAL** MCQ single answer **(male/female)** Gender confirmation
- **(AQ7) FACTUAL** Drop-down selection list **(6 options)** Age bracket identification
- **(AQ8) FACTUAL** Drop-down selection list **(all countries in the world)** Confirmation of nationality (most closely associated with) and free-text 'other 'option for added flexibility
- **(AQ9) FACTUAL** 4 sub items in a grid format **(A-D)** MCQ single answer **(yes/unsure/no)** investigating EAP education and assessment involvement
- **(AQ10) FACTUAL** Drop-down selection list **(all countries in the world)** identification of country where the participant has most recently been involved with EAP assessment and free-text 'other 'option for added flexibility
- **(AQ11) FACTUAL** Drop-down selection list **(6 different types of educational institution)** Description of type of Institution where most recently involved with EAP assessment and free-text 'other 'option for added flexibility
- **(AQ12) FACTUAL** Drop-down selection list **(5 different time spans from 0-23 months to over 20 years)** Confirmation of time elapsed since first started working in EAP

Save & Continue to Next Page

(Page 4)

Training and Skills in EAP training and assessment

- **(AQ13) ATTITUDINAL** 10 sub-items **(A-J)** in a grid format Likert scale MCQ single answer **(strongly agree/ agree /neither agree nor disagree / disagree/ strongly disagree)** investigating participant training and skill in EAP testing and assessment
- **(AQ14) BEHAVIOURAL** 13 sub-items **(A-M)** in a grid format Likert scale MCQ single answer **(very highly skilled/ highly skilled /aver\average skilled / low skilled/ very low skilled)** inviting participants to rate their skills in the processes and activities associated with EAP testing and assessment
- **(AQ15) ATTITUDINAL/BEHAVIOURAL** 1 free-text multiple line text box inviting participants to add additional comments about training and skills in EAP assessment and inviting the clarification of any responses given to closed questions

Save & Continue to Next Page

(Page 5)

Views on EAP testing and Assessment Design processes

- **(AQ16) ATTITUDINAL** 10 sub-items **(A-J)** in a grid format Likert scale MCQ single answer **(strongly agree/ agree /neither agree nor disagree / disagree/ strongly disagree)** inviting participants to identify views on the EAP test and assessment design process
- **(AQ17) BEHAVIOURAL** 8 sub-items **(A-H)** in a grid format, each with a drop-down selection list (9 options). Each of the 8 items represents a stage in the EAP test design process and participants are asked to identify their preferred sequence.
- **(AQ18) ATTITUDINAL/BEHAVIOURAL** 1 free-text multiple line text box inviting participants to add additional comments about training and skills in EAP assessment and inviting the clarification of any responses given to closed questions

Save & Continue to Next Page

(Page 6)

Views on validity and reliability in EAP testing and assessment

- **(AQ19) ATTITUDINAL** 10 sub-items **(A-J)** in a grid format Likert scale MCQ single answer **(strongly agree/ agree /neither agree nor disagree / disagree/ strongly disagree)** asking participants to comment on their attitude towards situations concerning validity and reliability in EAP testing and assessment
- **(AQ20) ATTITUDINAL** 8 sub items in a grid format **(A-H)** MCQ single answer **(yes/unsure/no)** investigating participant views about actions which are likely to enhance validity
- **(AQ21) ATTITUDINAL/BEHAVIOURAL** 1 free-text multiple line text box inviting participants to add additional comments about validity and reliability and inviting the clarification of any responses given to closed questions

Save & Continue to Next Page

(Page 7)

Views on analysis and interpretation of tests and assessment results

- **(AQ22) ATTITUDINAL** 10 sub-items **(A-J)** in a grid format Likert scale MCQ single answer **(strongly agree/ agree /neither agree nor disagree / disagree/ strongly disagree)** asking participants to comment on their attitudes and experience of situations concerning test and assessment analysis and interpretation
- **(AQ23) BEHAVIOURAL** 10 sub-items **(A-J)** in a grid format Likert scale MCQ single answer **(strongly agree/ agree /neither agree nor disagree / disagree/ strongly disagree)** asking participants to rate their familiarity with methods of descriptive and inferential statistical analysis
- **(AQ24) ATTITUDINAL/BEHAVIOURAL** 1 free-text multiple line text box inviting participants to add additional comments about analysis and interpretation of EAP tests and assessment results and inviting the clarification of any responses given to closed questions

Save & Continue to Next Page

(Page 8)
Views on ethical considerations in EAP testing and assessment

- **(AQ25) ATTITUDINAL** 10 sub-items **(A-J)** in a grid format Likert scale MCQ single answer **(strongly agree/ agree /neither agree nor disagree / disagree/ strongly disagree)** asking participants to comment on their attitudes towards a series of ethical dilemmas associated with EAP testing and assessment
- **(AQ26) BEHAVIOURAL** 8 sub items in a grid format **(a-h)** MCQ single answer **(yes/unsure/no)** investigating participant ethical stance and concern for their own situation and ethical responsibilities in EAP language testing and assessment.
- **(AQ27) ATTITUDINAL/BEHAVIOURAL** 1 free-text multiple line text box inviting participants to add additional comments about ethical concerns in EAP testing and assessment and inviting the clarification of any responses given to closed questions

Save & Continue to Next Page

(Page 9)
Invitation to join a prize draw and to participate in interview stage

- **(AQ28) FACTUAL** 1 free-text single line text box inviting participants to add their email address for inclusion in prize draw
- **(AQ28) FACTUAL** 1 free-text single line text box inviting participants to add their email address and volunteer to participate in the interview stage

Save & Continue to Next Page

(Page 10)

Thank you and confirmation of data submission

- Thank you message and confirmation of data submission
- Invitation to contact the researcher by email with any related queries
- Reminder of future contact for participants who opted in to the interview stage

Table 30 Key areas of EAP Assessment Literacy linked to items/sections within the research questionnaire

Key areas of EAP Assessment Literacy identified for exploration in this study	Coverage via questionnaire items/sections
Training, skills, strengths and weaknesses in EAP testing and assessment	AQ13, AQ14, AQ15
Approaches to EAP testing and assessment design	AQ16, AQ17, AQ18
Views on/ experience of validity and reliability in EAP testing and assessment	AQ19, AQ20, AQ21
Practices involving analysis and interpretation of the results and scores of EAP tests and assessments	AQ22, AQ23, AQ24
Consideration of ethics in EAP testing and assessment	AQ25, AQ26, AQ27
Influence of research and other resources for the purpose of EAP testing and assessment on EAP teacher Assessment practices	AQ13 (In particular, AQ13d)

Fulcher (2012a) describes one of the limitations of previous investigations into Assessment Literacy by referring to the fact that the majority of studies have collected data via closed response sets in questionnaires or surveys. It was also noted by Fulcher (2012a) that closed response items can cause respondents to recycle the language which they have encountered when subsequently presented with more open responses in the same survey. For similar reasons it was decided that for the purposes of this study, the number of open items would be kept to a minimum and that more open questions would be dealt with through a separate interview procedure (Fulcher, 2012a, p.118).

While both Brown, 2001 and Fulcher, 2012a warn of difficulties associated with survey structure, Fulcher particularly recommends that survey item ordering should be added to Brown's list of disadvantages, due to the influence on open questions which information presented in front-placed closed questions can have. Nevertheless, for the purpose of this questionnaire, the ordering was not altered given that only one very open and generally focused open response option was included after each set of closed response questions. Furthermore, for the same reasons as listed by theorists interested in the role played by schema in reading comprehension, a thematic link or introduction to the topic in this survey could be considered purposeful as it may provide respondents with a contextual focus for the particular item in question (Carrell et al., 1988; Wallace, 1997, p.371; Widdowson, 1983), provided that the survey is appropriately compiled and

the open questions are carefully worded, the order should not necessarily be considered compromising or leading.

3.4.5 Sampling the target population – identifying an appropriate source of questionnaire respondents

Questionnaire respondents were attracted through sending an email request with an embedded link to the web-based questionnaire provided through the BOS survey system (University of Bristol, 2012). The EAP-related distribution lists which were utilised for this purpose are as follows:

- The UK-based email-discussion list provided by BALEAP (2012b)
- The USA-based online-discussion platform offered through the TESOL International Association (TESOL, 2012)
- The Australian-based email forum provided through the Association for Academic Language and Learning (AALL, 2012)
- The internationally-focused UK-based email-discussion list for practitioners in English for Specific purposes, provided through the IATEFL ESP Special Interest Group (IATEFL ESP SIG2012)

In the cases of IATEFL and BALEAP, the researcher was already a member of these organisations and was able to send an email message with the link to the survey at no extra cost. The Australian-based AALL organisation agreed to circulate the message on behalf of the researcher with no requirement to join the forum; however for the USA-based TESOL International Association it was necessary to pay an annual subscription in order to gain the right to circulate a message on the distribution list.

In sending the email requests, every effort was taken to ensure that the audiences were appropriate so that the questionnaire would be received by recipients to whom it would be relevant. This often involved the distribution of links to the questionnaire via special interest groups within larger organisational structures, such as the English for Specific Purposes division of IATEFL. The use of professional discussion fora as a means of distribution also guarded against selective bias (Gillham, 2000c) or the targeting of respondents with whom the researcher is acquainted. Although anonymity has been protected, participants were given the option to provide email contact details to allow contact for participation in the interview research stage. At no point was this information linked in a manner which would risk bias or disclosure of personal information during the analysis or reporting of data. Further details relating to the filtration of the total population of respondents which was undertaken in order to identify a suitable sample for further scrutiny will be discussed further in Chapter 4.

3.4.6 Questionnaire piloting

In advance of administering the questionnaire which was used for data collection in this project, a piloting and trialling process was undertaken in order to anticipate and consider possible pitfalls or shortcomings which may occur before the survey was used with the wider population of respondents (Gorard, 2001, p.111). The process of piloting involved two stages in order to review the initial questionnaire, to revise it and then to test it again so that further revisions could be made in before the final usage.

The first phase involved the piloting of a paper-based version of the questionnaire with two colleagues and two friends in order to identify possible problems associated with the questionnaire from both the lay and expert perspectives. Each participant was asked to complete the

questionnaire as if they were completing it as respondent from the target population but also to highlight any items with problematic features or which they found frustrating.

After each respondent had completed the questionnaire, a separate discussion was held with each pilot respondent during which each of the items and sections of the questionnaire were discussed. This resulted in the alteration of certain features. As a result of this phase, the following issues were noted:

- Ambiguous or overly technical language was revised and clarified
- The sequencing or location of questions was changed in order to improve the representative nature of constructs in Likert scale groupings
- Leading questions were removed
- The indication of the amount of time required for survey completion was increased in order to be more representative and accurate
- A series of repetitive or overlapping questions were removed or rephrased.

Although there was some adverse feedback at this stage regarding the time-consuming nature of the questionnaire, it was considered unavoidable given the researcher's wish to retain both positively and negatively phrased versions of each question in an attempt to minimise extreme response bias and response set acquiescent bias (Bryman, 2008, p.658). This phenomenon describes the situation which arises when respondents display a tendency to respond in a certain way regardless of the item's content. This can arise as a result of over use of Likert items where respondents find the format too repetitive. After discussion with the pilot respondents, this seemed to be the case.

In the second stage of the piloting process, two additional colleagues were asked to complete the revised survey in its online format and to suggest any further modifications. This phase tested the reception of the online-format as well as the nature of the questions and their organisation.

A number of additional corrections were noted at this stage, including a number of items which were left optional which needed to be made mandatory in order to avoid omission by some respondents. However one of the major problems which required rectification at this point was associated with the on-going concern regarding response-set acquiescence (De Vaus, 2002, p.107).

Despite the researcher's initial preference to include a version of each question, phrased both negatively and positively, it became apparent for a number of reasons that this was not a successful strategy. Whilst the possibility may exist to minimise response bias or response set acquiescence by varying the question wording and using negative phrasing, other guidance on good practice for survey development suggests that negative questions should be avoided due to the likelihood of misinterpretation and confusion caused by negative statements (Brown, 2001, p.47). Furthermore, it was apparent that rather than reducing response set acquiescence the inclusion of both positively and negatively phrased items doubled the length of the Likert scales included and therefore caused the very acquiescence that the researcher was seeking to avoid. This conclusion is corroborated by Sauro (2011) and Sauro and Lewis (2011) whose research indicates that until the net benefits of alternating item wording are proven, it remains better to retain an approach which uses positive wording.

3.4.7 Considerations and good practice in conducting interview research

According to Gillham (2000c, p.12), any research which aims to achieve an understanding of people in a real-world context will require some interview material. As with a number of other

research methods discussed, interviews can be categorised along a scale (Nunan, 1992, p.149) which can range from the unstructured, and less directive method (Bryman, 2008, p.196) towards the more structured alternative (Drever, ibid, p.1). For a semi-structured interview often the main items are open in nature whereas structured interviews frequently use more closed questions (Gillham, 2000c, p.41; 2005, p.82). The purest qualitative form of interview is considered to be the informal qualitative interview (Tashakkori & Teddlie, 2003, p.305). Ethnographic interviews of this nature aim to provide detailed, thick descriptions of a particular culture or section of society (Dörnyei, 2007, p.130). Typically the process involves researcher curiosity and the development of participant trust whilst maintaining research unobtrusiveness (Cohen et al., 2007, p.350).

Usually, semi-structured interviews contain some specific, or closed questions but remain flexible (Denscombe, 2007, p.176) allowing for elaboration and modification in the formulation of questions and answers so that there is scope for the unexpected to emerge and to be captured (Gillham, 2005, p.82; Seliger & Shohamy, 1989, p.167). In the main, research interviews of this type include a mixture of main questions, follow-up questions with the option to use probes where necessary (Rubin & Rubin, 2005, p.152).

A researcher may choose to use an interview, when it is necessary to gain insights into people's views, feelings and experiences (Denscombe, 2007, p.174). This method also provides a mechanism for collecting data from field specialists such as subject specific informants (Douglas, 2000, p.33). Interviews can also be used together with other research methods (Cohen et al., 2007, p.268) in order to follow up unexpected results, validate alternative methods and investigate deeper into respondents' motivation and rationale for responding in a particular way (Kerlinger & Lee, 2000, p.693).

With this in mind, in many respects, the research interview has been selected for use in this project to attempt to address some of the shortcomings as identified in the use of questionnaires. For this reason, amongst others, interviews are among the most frequent qualitative research tools (Mason, 2002, p.62). It is also perhaps unsurprising, given the interpersonal nature of teaching and learning, that talking to students and/or their instructors is a logical recourse (Drever, 2003, p.1) for this research into EAP teacher Assessment Literacy. Some of the key strengths and weaknesses of this form of data collection are listed below in Figure 16:

Figure 16 Strengths and weaknesses of interview usage for this research

Strengths and means of harnessing	Weaknesses and means of mitigation
• **Flexible structure** Adaptation and deviation is possible when the unexpected emerges. Every effort needs to be taken to keep on track with research questions. • **Extensive means of asking** The interviewer has used different questioning methods to solicit views and opinions. Leading questions have been avoided. • **Allows for probing** The Interviewer has been able to follow-up or ask for further clarification where deemed relevant. • **Easy to arrange** Time and a suitable conditions have been the key resources and this has been be accompanied by planning and piloting. • **Allows unexpected discovery** The semi-structured interviews has allowed for the unplanned to emerge and to be followed up. However, due care has had to be taken to avoid being side-tracked. • **Can elicit rich in-depth descriptions** The interviews have been useful data collection method to support the use of questionnaires where probing was not possible. The interviews were time-consuming to transcribe. • **Useful for exploration and confirmation (triangulation)** The patterns in data collected from the interviews can be compared with themes which have emerged from the questionnaire. • **Measurement validity is strong if interview is well constructed and tested** Questions have had to be carefully chosen, formulated and piloted. • **Can be used in conjunction with other methods** The interview is commonly considered complementary, as a follow-on stage from a questionnaire.	**Time consuming to analyse** (especially open questions) Has required the use of analysis grids where both categories and units of analysis are then identified.(Cohen et al., 2007, 164-165). • **Requires transcription** This stage has been time-consuming, as experienced in the pilot stages of this project. Reference has been made to transcription methods, as suggested by Jefferson (2004, in Lerner, 2004, p.13-31) In addition to, Potter and Hepburn (2009) and Van Lier (1988, p.238-244). The appropriate use of technology has also be considered (Evers, 2011). • **Results may be unrepresentative** As the number of interviewees is limited the risk of the sample being unrepresentative of the wider population has been minimised by involving a series of interviewees from across the sector with experience of working in different context. • **Questionable reliability** As interviews are time-consuming, a large number of interviews have not been possible; therefore this method has not been used in isolation but rather in conjunction with a questionnaire tool. Interviews have been used for corroborative purposes in this project, however the qualitative stance on the value of thorough exploration of local phenomenon also challenges a restricted view of generalisability (Duff in Chalhoub-Deville et al., 2006, p.74) • **Interviewer bias may affect interviewee** Due care has been taken to minimise bias, not to prompt interviewees or to ask leading questions. • **Requires careful question construction and piloting.** Questions have been designed and trialled with the research questions and interview questions in mind in order complement questions used in the questionnaire.

(Adapted from Bynner & Stribley, 1979; Cohen et al., 2007; Denscombe, 2007; Dörnyei, 2007; Drever, 2003; Gillham, 2005; Kerlinger & Lee, 2000; Tashakkori & Teddlie, 2003)

3.4.8 Interview design

In order to commence the interview design process, the purpose of the interview was identified in conjunction with the research questions for the project and the features of the variables being explored (Nunan, 1992, p.151). Importantly, this involved a clear definition of the construct under investigation, as discussed within the Literature Review. This stage was essential in order to establish accurately the type of the data which needed to be collected. Mason (2002, p.48-51) provides a useful planning and preparation tool which allows the researcher to breakdown the main research questions into smaller elements; this information could then be used to help develop interview questions.

Decisions were made concerning the use and sequencing of questions as well as choices from the range of item structures which are available.

One of the questions which qualitative research associated with interviews needs to answer is, 'how many interviews need to be conducted?' Bryman (2012, p.425) draws attention to research conducted by Warren (2002, p.99) which indicates that for research to be published, the minimum number of interviews required appears to be between 20 and 30. Despite this suggestion, Bryman (ibid) lists other examples which dispute this fact at both ends of the spectrum. Occasions are cited when fewer interviews may be required, for instance when longer or intensive interview models are conducted. However, other researchers believe that a greater number of interviews is required in order to support convincing conclusions. This situation shows how views differ and how difficult this aspect of decision making related to the research model can be. It seems to be the case that the sample should be sufficiently large so as not to allow data saturation but not too large which could render analysis and interpretation too complex.

Given the resources available to the researcher in terms of time and accessibility to prospective interviewees, along with the aim of using interview findings in conjunction with the findings from the questionnaire, a target of 25 interviews was set. Due to this limited number of interviewees, and the single session of interviews which was undertaken, it was determined that the data elicited from this research method would be most valid and useful in terms of the qualitative insight which it would provide alongside the more quantitative data collected from respondents during questionnaire stage. In this manner it can be considered as a form of triangulation as it offers the opportunity to capture variant perspectives on a phenomenon based on different forms of data collection method (Heigham & Croker, 2009, p.11). Figure 12 on page 144 describes the triangulation process. With this in mind, it was decided that the use of more open questions than closed questions would be appropriate in order to allow interviewees the opportunity to express themselves more fully and richly (Denscombe, 2007, p.166).

A series of questions was developed which took into account the research questions for this project and the construct for Assessment Literacy as identified in the Literature Review. The set of interview items was then reviewed and condensed in order ensure that the interview encompassed only necessary items which were appropriate and distinctive from one another (Gillham, 2000c, p.21). Attention was paid to the ordering of questions so that respondents would feel comfortable with the content and tone from the start of the interview. Given the different range of experience amongst the research population, very technical terminology was also avoided, in recognition that not all interviewees would have knowledge of technical expressions associated with assessment (Dörnyei, 2007, p.137).

The full interview protocol which was used by the interviewer with interviewees is available in the appendix of this thesis in section in section 8.2 on page 437 . Figure 17 below provides an overview of the interview structure and objectives of each element and question item.

Table 31 shows how areas of EAP teacher Assessment Literacy, demonstrated through practice and research and summarised in the conclusion to Literature Review in Section 2.7, are connected to items in the interview protocol.

Figure 17 Interview structure and item features

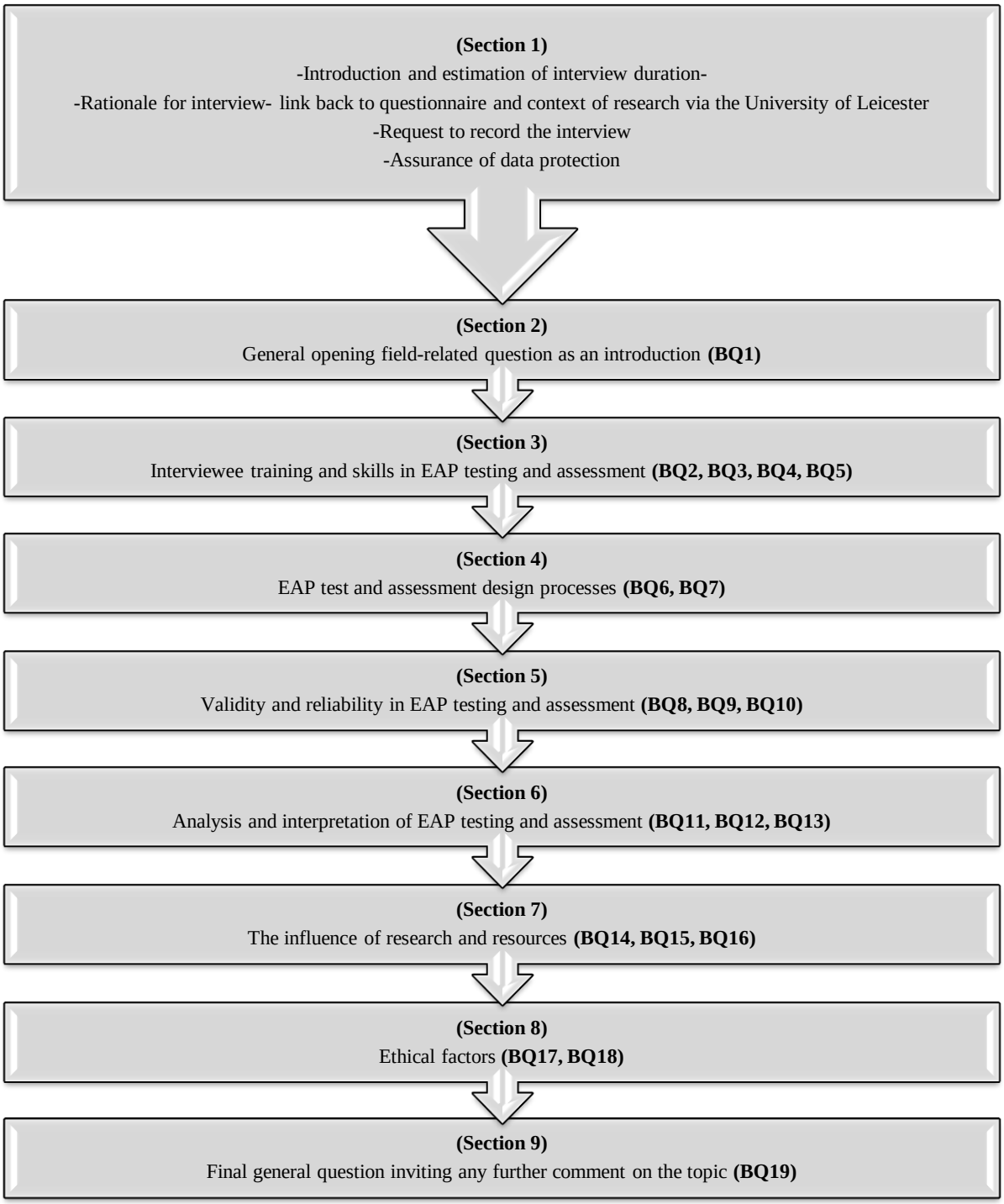


Table 31 Key areas of EAP Assessment Literacy linked to items/sections within the interview protocol

Key areas of EAP Assessment Literacy identified for exploration in this study	Coverage via interview items/sections
Training, skills, strengths and weaknesses in EAP testing and assessment	BQ2, BQ3, BQ4, BQ5
Approaches to EAP testing and assessment design	BQ6, BQ7
Views on/ experience of validity and reliability in EAP testing and assessment	BQ8, BQ9, BQ10
Practices involving analysis and interpretation of the results and scores of EAP tests and assessments	BQ11, BQ12, BQ13
Influence of research and other resources for the purpose of EAP testing and assessment on EAP teacher Assessment practices	BQ14, BQ15, BQ16
Consideration of ethics in EAP testing and assessment	BQ17, BQ18

3.4.9 The interview medium - face-to-face and Skype interviews

In considering the suitable situations in which the interviews could be undertaken, the researcher was presented with a number of options. The traditional method of interviewing involves the interviewer sitting opposite the interviewee and asking a series of questions whilst recording the responses either on a dictaphone or in note form (Bryman, 2008, p. 197). Figure 18 below provides a checklist used for satisfying the conditions under which a face-to face interview is appropriate, necessary or possible (Gillham, 2000c) along with annotations relating to this particular enquiry associated with EAP Assessment Literacy:

Figure 18 Indications of whether a face-to-face interview is appropriate, necessary or possible

No *if*…	**Yes if…**
Large numbers of people are involved	Small numbers of people are involved
A significant but relatively small number of interviewees will be sought for this project as a form of triangulation in conjunction with a questionnaire tool	
People are widely dispersed	People are accessible
Whilst the target population of EAP teachers is relatively widely dispersed with the country and potentially internationally, the use of face to face interviews in addition to telephone-type interviews has effectively made dispersal more manageable	
Most of the questions are closed	Most of the questions are open and require an extended response with prompts and probes
Given the intended use of the interview to complement the use of the questionnaire which adopts a more closed format, the questions in the interview are more open in nature.	
100% response rate is not necessary	Everyone and their view is key
The aim of the interviews is to collect as many views as possible from a range of representative members of the target population.	
The material is not subtle or sensitive	The material is sensitive in character so that trust is involved
Interviews have been identified as a suitable data collection method as the research in question involves the investigation of behaviours and attitudes which cannot be examined without some more direct and flexible questioning of members of the target population.	
You want to preserve anonymity	Anonymity is not an issue, though confidentiality might be
Confidentiality is considered more crucial in this particular research context as the issues under investigation are not so highly sensitive as to cause members of the target population to feel uncomfortable in participating in interviews.	
Breadth and representativeness of data are central	Depth of meaning is central
As breadth of representativeness has been the principal aim of the questionnaire the interviews have been utilised to add an element of further depth.	
Research aims are factual and summary in nature	Research aims require insight and understanding
The interviews have been designed to bring a deeper level of understanding through extended responses to open questions, which were not the main feature of the questionnaire.	

(Gillham, 2000c)

As a result of the considerations listed in Figure 18 above, it was decided that where possible a face to face model would be adopted for the semi-structured interviews which would be used in the research project.

However, in addition to more traditional forms of interviewing, it was also acknowledged that other options also exist for conducting interviews, including the use of telephone interviews or the use of computer media such as Skype, a Voice over Internet Protocol (VOIP).

Figure 19 below presents a number of the strengths and weaknesses of using telephone-type interviews along with further considerations relating to the particular research scenario associated with this thesis:

Figure 19 Strengths and weaknesses of using telephone-type interviews

Strengths and means of harnessing	Weaknesses and means of mitigation
• **Cheaper and quicker to administer given removal of the need to travel to interview people.** Although, where possible a number of face-to-face interviews have been conducted the use of telephone-type interview using Skype has enabled access to a wider population including interviewees overseas. • **Telephone interviews are easier to supervise, especially where there are a number of interviewers.** Although only one interviewer has been involved and only one interview has been conducted at any one time reference to interview questions and use of recording equipment has been relatively easier in the process of using an add-on software package which records directly from Skype. • **The relative remoteness of the interviewer could remove a source of bias through answers being affected by the interviewer's presence.** The researcher has not noted any obstacle presented by using telephone-type interviews in comparison to face-to-face options. In a number of situations it has been felt that the absence of a shared physical environment has allowed additional focus on the questions being posed.	• **People who do not have access to a telephone or other telecommunications equipment cannot provide their data in this manner.** Although, not all interview respondents who were invited by email to take part in the interview stage responded, it was possible to undertake the target number of interviews. • **The hearing impaired are likely to find this medium difficult to use** In the process of this research project, it was not noted by the interviewer or any interviewee that they experienced any difficulty in hearing. • **The length of the interview will be difficult to manage after more than 20-25 minutes.** Interviewees were given a clear indication of the minimum amount of time which would be required for the interview which was suggested to be approximately 20 minutes. Although in some cases interviews exceed this, it was only under circumstances where interviews appeared comfortable to continue. • **Response rates to telephone interviews seem to be slightly lower than face-to-face alternatives.** A total of 16 questionnaire respondents were invited to respond in order to conduct the 12 telephone-type interviews which were undertaken as part of the total number of 25 interviews • **Telephone interviewers are not able to record non-verbal cues such as facial expression.** The absence of non-verbal cues was not noted as debilitating as instances when it was required to note such features did not arise in the face to face interviews. Visual aids cannot easily be employed This research context did not require the use of visual aids. • **Some evidence suggests that quality of data may suffer through 'don't know' responses and relative lack of engagement.** In nearly all cases respondents attempted to answer questions posed fully and where further details were required additional probes were used. There was no indication that response duration was curtailed by the interview format.

(Bryman, 2008, p.198-199; Gillham, 2000c, p.12; Gwartney, 2007, p.15-16)

For the purpose of the interview research conducted for this project, it was finally decided that both telephone interviews and face-to-face interviews would be conducted in order to maximise the strengths and weaknesses associated with the different contexts and media involved. This approach was also adopted in order to permit access to a sufficiently large number of interviewees based in a range of different geographical locations. A number of face-to-face interviews were held at a conference, others were held on location at venues accessible to the researcher and the remaining interviews were conducted via Skype-to-Skype and Skype-to-telephone protocols.

3.4.10 Sampling the target population – identifying an appropriate source of interviewees

Participants for the study have been identified and invited to contribute based on volunteers who provided contact details during the initial questionnaire stage. An appropriate grouping of interviewees was determined through stratification in line with the characteristics of the questionnaire population. Further details relating to this are provided in Chapter 4.

As indicated in Figure 15, on page 9 of the questionnaire, questionnaire respondents were given the opportunity to opt in to the interview stage by providing their email address. Whilst it is true that this risked exposing the identity of participants, respondents were assured that at no point would their email address be divulged in a manner which would reveal their personal details or institutional affiliation. In addition, interview participants were selected at random from the overall list of respondents who provided their email contact details.

Interview times were scheduled via email, either to take place face-to-face in the free time surrounding an academic conference or to arrange a suitably convenient time for a Skype-conversation. Background guidance and explanatory notes were sent to prospective participants

along with, timing indications and assurance of anonymity. Prospective participants were also reminded of the research objectives and confirmation of ethical approval, full details of which were provided during the questionnaire stage. Based on the differing contexts presented by telephone-style interviews and face-to-face interviews consideration was given to the building of an appropriate rapport with respondents. This required an additional amount of introduction and contextualisation, related to the research project at the start of the interview in the case of the sessions undertaken via Skype, given the absence of visual cues (Bryman, 2008). Due care was taken however to ensure sufficient balance so that interaction with respondents remained sufficiently distanced to avoid any bias or leading comments.

It was decided that a semi-structured interview would be the best approach to follow in order to allow the inclusion of items specifically focused on key aspects of Assessment Literacy, as identified in the literature review for this thesis. The interview design model remained flexible (Denscombe, 2007, p.176) so that, where necessary, questions could be adjusted according to contextual factors and unexpected data could be recorded and further clarification sought (Gillham, 2005, p.82; Seliger & Shohamy, 1989, p.167).

In accordance with the protocols for semi-structured interviewing, in the design of questions, care was taken to avoid active solicitation of answers, whilst maintaining a free and open discussion which did not diverge from the topic of interest (Kerlinger & Lee, 2000, p.702), namely Assessment Literacy.

The researcher took on the role of interviewer. With further regard to the reliability and validity of data collected through the interview process, consideration was given to the possibility of using more than one interviewer, however the cost implications along with other accounts suggesting the

limited added value which this was likely to achieve in the context of more structured types of interview (Bryman, 2008, p.197), resulted in the decision to retain just one interviewer.

The full interview transcript which was used by the interviewer with interviewees is available in the appendix of this thesis in section in section 8.2 on page 437. Figure 17 on page 168 provides an overview of the interview structure and objectives of each element and question item.

3.4.11 Interview piloting

Through piloting or trialling, the interview method was tested, modified and improved in advance of its application to the full project (Seliger & Shohamy, 1989, p.195). Piloting involves using the interview schedule with one or more members of the representative group being researched in a form of dress-rehearsal (Gillham, 2000c, p.55). The aim is to identify defects which can be removed before usage with the main group of interviewees. The procedures below indicate the steps which have been taken in order to design and trial the interview protocol which has been employed for this research.

Once again, the trialling process was undertaken with the assistance of two colleagues, one internal and one external to the researcher's place of work. Both individuals were considered to represent the target population for research and indeed they were selected based on different levels of experience in the field. It was anticipated that that this manner of preparation could enhance the reliability of the interviews (Cohen et al., 2007, p.151). It should also be noted that the individuals involved in trialling were selected on the basis that they would not participate in the main series of interviews. One interview was conducted using Skype-to-telephone protocol and the other took the form of a face-to-face context. In both situations the same set of questions was trialled.

Suggestions for revision and amendment involved advice concerning question wording; sequencing and question suitability, as well as question groupings. Gillham (2000c, p.21-22)

explains that this form of trialling allows the researcher to gain a feel for the interview process and its direction. It allows focus on faulty or successful questions and highlights areas which require additional thought.

Through this process it was identified that an initial general introductory question should be added to the questionnaire at the start of Section 2 of the interview, in order to allow the open expression of participants' views before being asked more specific questions. A similar question was also located at the end of the interview, in Section 9, in order to allow participants to contribute any other comments that they may have on the topic of EAP assessment, should they not feel they had had sufficient opportunity through the structured questions provided.

The piloting process also indicated that the duration of the interview was too long and, as a result, the decision was taken to combine questions on validity and reliability into one single section which is now Section 5. These areas can be considered to contribute to the wider context of construct validity, although the term construct validity was not used in the interview as it was anticipated that there would some unfamiliarity with this term amongst some interviewees.

In both Sections 4 and 7 of the interview, it was decided that two questions should be conflated in order to avoid repetition, as both pilot respondents suggested that an answer to one particular question also solicited the answer to a subsequent question.

It was noted that in Section 8 of the interview, focusing on matters connected to the ethics in assessment, both respondents in the pilot required further explanation in order to understand the nature of information solicited by the question. As as a result the item was rephrased. It was noted however that further subtle and unbiased context may still be required during the actual, interview

as it was considered that any question which overtly asked, do you have any 'ethical concerns' may be leading in nature.

3.4.12 Data recording and transcription

As the research in this project has involved the analysis of interview data, it was considered necessary to transcribe the information collected through that process. This is due to the fact that real conversations are commonly fractured, hesitant and ungrammatical (Giddens, 2001) which requires closer examination in written form in order to ensure that the true meaning has been accurately identified.

The principal reference point for the purpose of recording data in this research project remains the transcription method suggested by Jefferson (2004, in Lerner, 2004, p.13-31). Additional options were also considered through reference to Duranti (1997) Denscombe (2007), Potter (2009), Van Lier (1988 p.238-244). During this process of consideration, the researcher took into account the fact that the type of the research project itself has considerable bearing on the extent to which broad or narrow transcription method should be adopted (Gee, 2011). Given the training and experience gained through earlier stages of the research before the thesis stage, it was determined that a relatively broad application of Jeffersonian transcription would still be suitable for the transcription of interviews conducted for the main thesis. Further reference to the options described by Evers (2011) also assisted in finalising the transcription method.

Questions which could arise when deciding how to transcribe a stretch of spoken text (Van Lier, 1988, p.238-244) may include the following:

- How much information should be included?
- How accurate should the transcription be?

- How much non-verbal communication should be transcribed?
- Who should do the transcribing?

With regard to the transcription undertaken for this project, the transcription aimed to reflect exactly what was said by the interviewer and the interviewee (Brown, 2001, p.100), however very few non-verbal features were transcribed as the nature of discussions did not appear to render these features particularly significant.

The researcher decided that it would be necessary to undertake the transcriptions himself, despite the time-consuming nature of the work involved, in order to be able to start to examine the data with closer proximity and to start to identify emergent trends (Bryman, 2008, p.456).

In order to make the best use of available technology (Brown, 2001, p.100) an electronic foot pedal was purchased along with a digital recording adaptor for use with iPod©. This equipment was used with transcription software which can be obtained as freeware downloaded from the internet. The information below provides a description of the equipment used:

- MP3 Recording tool - iTalk Pro© (Griffin Technology, 2009)
- MP3 player/recorder - iPod classic (Apple, 2009)
- Transcription audio Playback software - (NCH Software, 2009)
- Transcription Foot Pedal control - Phillips LHF 2310 (Phillips, 2009)

The transcripts from each of the 25 interviews conducted are available for download via hyperlinks which are provided in the appendix of this thesis in section 8.3 on page 439.

3.4.13 Section conclusion

This section of the study has examined more closely the process of designing research tools specific to this research project and the research questions which guide enquiry. Good practice and key considerations for the development of questionnaire and research tools have been explored in the context of research into EAP Assessment Literacy. In addition, approaches to sampling the target population have been outlined and piloting procedures have been discussed. This has has assisted in adjusting the data collection tools. In turn, this has also contributed to enhancing the reliability and validity of inferences drawn from the data collected.

3.5 Selecting and implementing data analysis tools

3.5.1 Section introduction

This part of the thesis seeks to examine the nature of the data collected in this study related to EAP teacher Assessment Literacy. This will then enable the selection of suitable methods for data analysis. Consideration will be given to the features of the data stemming from both the questionnaire and interview tools and how the data sets are best analyzed in the light of the research questions which have been identified in this study. Reference will also be made to the process of coding and identifying substantive points and categories within the data.

3.5.2 Data features, categories and measurement scales

In advance of considering which methods of analysis will be used to examine the data collected in this project, it is useful to consider the in-built categories and measurement objectives pertaining to the items used to collect data.

The interview items and open questions from the questionnaire have been designed in line with key themes associated with Assessment Literacy in EAP, as summarised at the end of Chapter 2.

This has led to the collection of open responses which would require qualitative analysis. In contrast, different kinds of quantitative measurement have been used in the questionnaire which requires quantitative analysis. Table 32 on page179, based on Brown (1988, p.20-23), presents a common range of quantitative measurement scales and examples which characterise their usage. As Table 32 suggests, it is nominal and ordinal scales which are most frequently employed in the questionnaire tool used in this research. Further insight into the design of individual items from both the questionnaire and interviews can be gleaned with reference to Figure 15 and Figure 17 on pages 155 and 168 respectively.

Table 32 Measurement scales (Brown, 1988)

Scale Type	Features	Example	Calculating central tendency/ distribution
Nominal	Provides names and categories only (no order implied)	Dichotomous: Gender- Male and female Non-dichotomous: Nationality- British, American, Korean	Mode only
Ordinal	Provides categories but also provides the ordering or ranking of those categories (no information about distance between categories)	Dichotomous: True/False Non-dichotomous: Ordinal scales include Likert scales which consist of multiple values, such as ‘strongly agree', ‘agree', ‘neither agree nor disagree’, ‘disagree' and ‘strongly disagree’	Mode or Median/ Range, Interquartile Range
Interval	Provides categories and ordering and shows distance between points in that ordering. Notably intervals have the same interpretation throughout (no absolute Zero)	Final examination scores in a test: If student ‘A’ scored 70, student ‘B’ scored 85 and student ‘C’ scored 95 in a test out of 100 this could be ranked in order of score in 1st, 2nd and 3rd place, thus representing an ordinal scale, however more information is given about the distance between each students’ attainment in addition to their ranking which constitutes an interval scale. There is no absolute Zero, for example the attainment of score of zero on a language test does not represent the complete absence of language ability.	Mode, Median and Mean / Range, Interquartile Range / Standard Deviation
Ratio	Provides the intervals between points in the ordering of certain categories but with more information because there is an interpretable zero and multiples of points along the scale make sense.	Electricity It is possible to have zero electricity in a house. If a 50 watt bulb is turned on and then an additional 100 watt bulb is illuminated, now three times the amount of electricity is being used. Ratio scales are not so applicable to the context of this study because, for example, it is difficult to say that someone has zero EAP Assessment Literacy skill or that (s)he is three times as skilled in assessment than another person.	Mode, median, mean or geometric mean (multiplying a set of figures and dividing by the number of figures multiplied, e.g. useful for calculating average rates of growth) / Range, Interquartile Range / Standard Deviation

Through understanding the nature of the data collected, along with its features and inherent categories, this can assist in the process of identifying methods for analysis, which will be discussed in Section 3.5.3.

3.5.3 Methods of analysing the data collected

For the purpose of analysing the data collected through the questionnaire and interview methods utilised in this research project, a series of tools are identified in Table 33 and Table 34.

Table 33 Methods to analyse quantitative closed-response questionnaire data

Tool	Purpose
Descriptive statistics of grouping and charts indicating frequency	To demonstrate the central tendency of responses to questionnaire items.
Chi Square tests	To identify differences between different sets of variables.
Cronbach's Alpha (reliability Coefficient)	To measure the internal reliability of items in a Likert scale which have been identified as representing a particular aspect of Assessment Literacy.
Factor Analysis	To determine the construct validity of Likert scale items by reducing a large number of variables through a questionnaire to a smaller number of factors.

Table 34 Methods to analyse qualitative open-response interview and questionnaire data

Tool	Purpose
Content analysis & pattern coding	To identify and organise patterns in qualitative data.
Evaluation of interactional profiles	To consider circumstances and key features of each interview and interaction between the interviewer and interviewee.

The rationale for selecting the methods of analysis, as described in Table 33, for the purpose of analysing quantitative closed-response questionnaire data, will now be presented.

Descriptive statistics of grouping and charts indicating frequency will be utilised in order to identify central tendencies which emerge from individual questionnaire items and Likert scales. Chi-Square has been identified as an analysis method which can be used to identify differences between different sets of variables. More specifically, it has been deemed appropriate as it is an inferential statistic which is applicable for use with ordinal data such as that gained from the main DVAS items in this study.

Cronbach's Alpha will be used to gauge the internal reliability of items in a Likert scale which have been arranged with the expectation that they represent a particular aspect of Assessment Literacy (Dörnyei, 2007, p.206-207). This helps to identify whether in most cases, respondents who rated Question A highly also rated Question B highly. In addition, Cronbach's Alpha can identify whether those who rated Question A as low also tended also to rate Question B as low. This type of correlation indicates a reliable representation of a construct in a survey. However, if in response to Question A and Question C, respondents gave notably divergent ratings, this absence of correlation may suggest that the items do not measure the same construct and therefore C should be removed or rephrased.

Factor analysis can be used to examine the construct validity of Likert scale items through the process of reducing a large number of variables through a questionnaire to a smaller number of factors (Bachman, 2004, p.257-293; Fulcher & Davidson, 2007, p.184). In the case of this research project, Exploratory Factor Analysis (EFA) has been identified as preferable to Confirmatory Factor Analysis (CFA). Whilst CFA is likely to show the extent to which the hypothesised model or construct fits or more likely doesn't fit, EFA is designed to help make sense of the data by identifying constructs.

The selection of methods for the purpose of analysing qualitative open-response interview and questionnaire items, as outlined in Table 33, will now be explained. Content Analysis has been identified as an appropriate analysis method as it examines and organises patterns in data (Tashakkori & Teddlie, 2003, p.314). For the purpose of this research project it will be used with transcribed interviews in order to identify meaning or trends embedded within qualitative responses. In qualitative Content Analysis categories of meaning surface through close

examination and, according to Denscombe (2007, p.237), Content Analysis can reveal otherwise invisible features within texts.

Reference has also been made to guidance on pattern coding from Miles and Huberman (1994, p.69), given that pattern coding can:

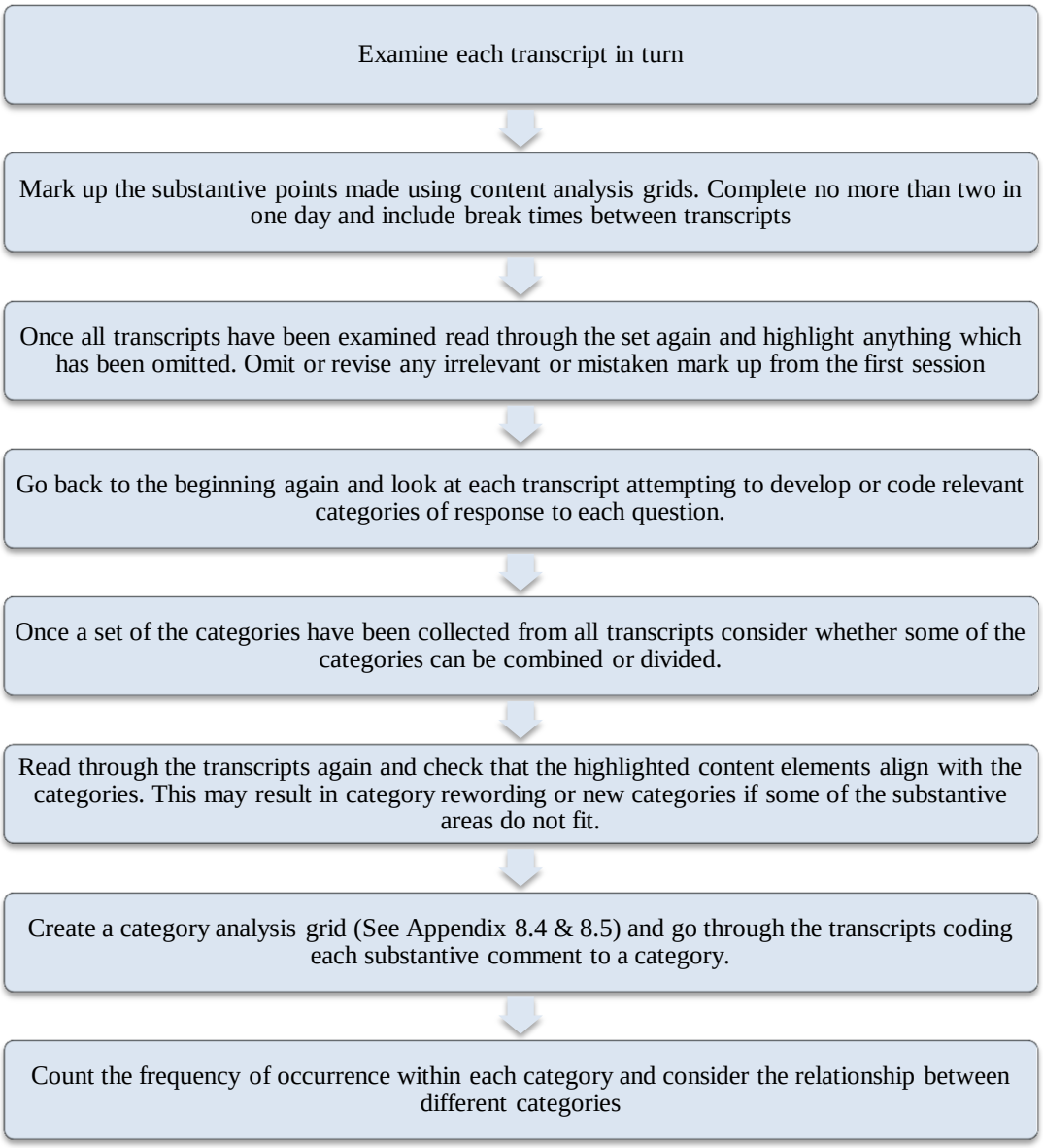
- enable the reduction of large amounts of data into smaller analytic units
- place the researcher in the analysis process during data collection
- help the researcher to develop a cognitive map for understanding incidents and interactions
- allow common themes to surface

The importance of coding will also be discussed in further depth in Section 3.5.4.

Before the thesis stage of this project, a means of Content Analysis was piloted for use with the transcribed data from a set of short interviews. This involved a series of procedures as suggested by Gillham (2000c, p.63-66) and Denscombe (2007, p.237). This method involved the use of analysis grids which assisted in the identification of categories and trends in shared meaning.
As this process was successful this structure has been reviewed and then reapplied for the final thesis.

Figure 20 on page 183, has also been included in order to highlight a model process for content analysis which has driven the researcher's approach.

Figure 20 Content analysis of transcribed interviews Gillham (2000c, p.63-66) and (Denscombe, 2007, p.237)



Finally, interview interactional profiles will be evaluated in order to consider the descriptions of features associated with interviewee interactions with the interviewer, in order to consider key features of each interviewee's case and characteristics of the interaction which may have affected the interviewee's particular contribution (Drever & Scottish Council for Research in Education. 2003, p.72 and Heigham & Croker, 2009, p.19). For this process a series of grids displaying key facts relating to each interview will be examined.

3.5.4 Coding and determining substantive points,

As described by Heigham and Croker (2009, p.308), coding refers to the phase of data analysis whereby researchers attempt to interpret the data which has been collected by looking through it methodically, grouping and clustering related substantive concepts and labelling them. The role of coding in Content Analysis is part of the process of disclosing aspects of what is being communicated (Denscombe, 2007, p.237).

For the purpose of quantitative data analysis, before the advent of advanced software, researchers may have had to put together a codebook, however given the facilities available through packages such as SPSS, the process of coding can be undertaken and recorded through electronic means. Nevertheless, despite the efficiencies provided by software, the coding process for quantitative data still involves certain manual stages, including creating the data file, defining coding frames and keying in the data (Dörnyei, 2007, p.200). Although the questionnaire data has been collected via electronic survey software which has not required the manual entry of survey responses, it has still been necessary to use SPSS to code response in numerical form and in some cases to recode negatively phrased likert scale responses before the use of Cronbach's Alpha and Factor Analysis.

The process of coding can assist researchers in familiarising themselves with the data, as well as improving data accessibility. During the data analysis for this research project, peer checking (Dörnyei, 2007, p.61) has been used as a mechanism to enhance interpretive validity. This has involved gaining the assistance of appropriately experienced colleagues in the process of developing and testing coding schemes. In this manner it has been possible to demonstrate that the codes and categories selected have been identified objectively and logically (Denscombe, 2007, p.302) through corroboration by another coder with relevant experience. This activity, combined with the piloting phases of data collection methods, as described in Section 3.4.6 and 3.4.11 can

be considered to contribute to the reliability of subsequent interpretations drawn from the data which has been collected.

For the purposes of this research into EAP teacher Assessment Literacy, data coding has been required in a number of different cases of data analysis. This has involved the coding of data as part of the content analysis of open-response interview and questionnaire data (See Table 34 in Section 3.5.3) and the identification of suitable labels for factors or categories which have emerged through the process of the factor analysis of quantitative data, drawn from questionnaire DVAS items and Likert scales.

Figure 21 and Figure 22 below describe the coding procedures involved in this research project where a seconder coder has been utilised to enhance validity:

Figure 21 Second coder usage in questionnaire data analysis

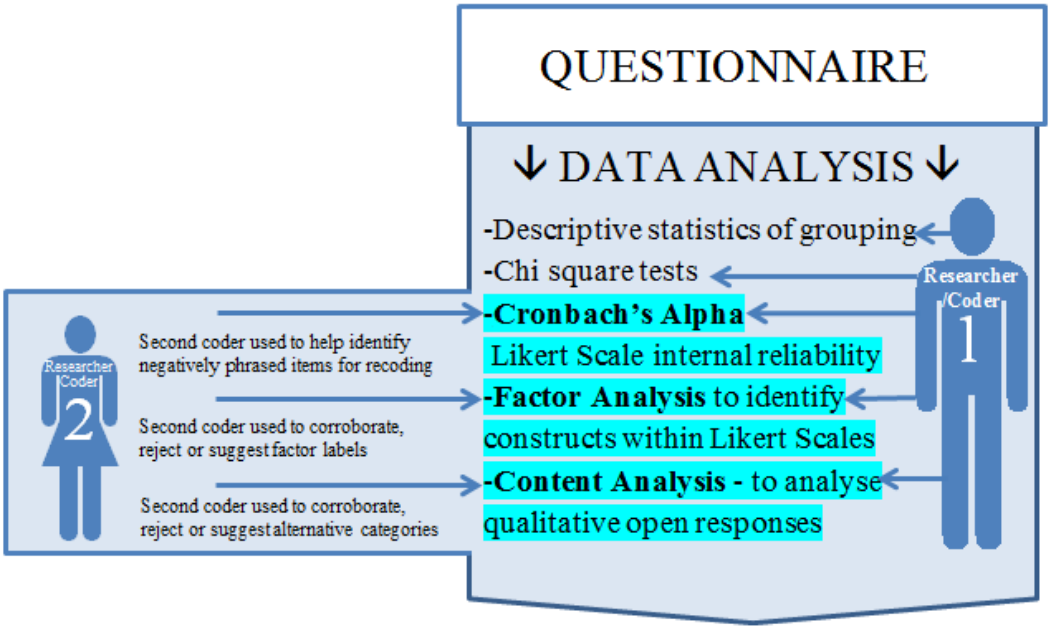
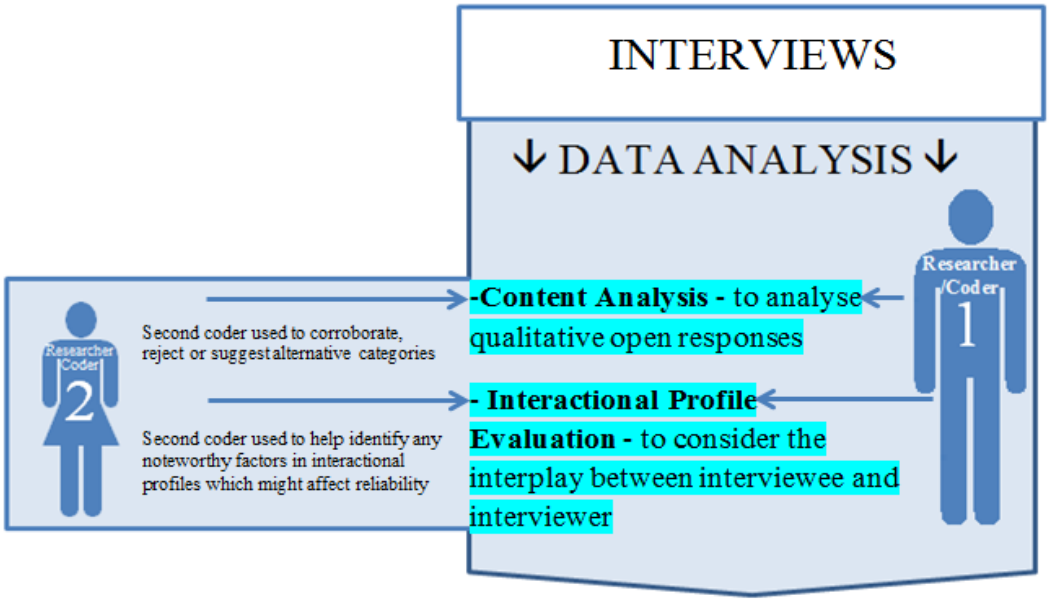


Figure 22 Second coder usage in interview data analysis



3.5.5 Section conclusion

In this section of the thesis, the process of selecting and using data analysis tools, which meet the requirements of the data collected in this project have been discussed. The objective of each proposed analysis method has been introduced and aspects of the process associated with coding and the determination of substantive points have been presented. A key element of the process of enhancing reliability has been outlined through reference to the usage of second coders.

3.6 Chapter conclusion

During the course of this chapter, the range of paradigms which might be suited to research in the field of EAP Assessment Literacy have been explored and the pragmatic approach has emerged as the most appropriate choice for this study. The range of quantitative and qualitative research methods, which might be applicable to the research questions which drive this enquiry, have also been investigated and, as a result, a mixed methods approach, using both a questionnaire and an

interview protocol, has been adopted. The design of data collection tools has been discussed along with piloting and sampling procedures. Finally, methods of data analysis and data coding have been described. The culmination of this process now leads on suitably to the next chapter of the thesis where the research findings will be introduced.

CHAPTER 4 FINDINGS

4.1 Chapter introduction

This chapter of the thesis will introduce and present the data which has been collected through the preceding stages of research and will describe the findings in relation to the research questions and hypotheses which have driven investigation. Interpretation and analysis of the findings will be undertaken in Chapter 5.

In order to organise this section and to restate the aims and objectives which the thesis has focused on, it is useful at this point to refer back to the research questions and hypothesis which relate to this particular research agenda in EAP teacher Assessment Literacy. In addition, the research questions which support this study were initially described in Table 1 which has been repeated below in Table 35.

Table 35 Thesis Research Questions (revisited)

Research Questions
• To what extent do EAP teacher views on EAP testing and assessment practices reflect language testing research and practices which comprise Assessment Literacy?
• How can EAP Assessment Literacy be sustained or enhanced?

The hypothesis which was originally articulated in Chapter 1 for the purpose of this research project is:

EAP teachers who are involved in EAP testing and assessment have certain identifiable development requirements with regard to their knowledge and ability to implement assessment good practice and recommendations stemming from research.

As explained in Section 1.2, in order to explore the above theory, the null version (H_o) of this hypothesis also needs to be considered, in order to investigate if it can be rejected and

consequently if support is discernible for the original hypothesis itself. The null hypothesis is described follows:

EAP teachers who are involved in EAP testing and assessment do not have identifiable development requirements with regard to their knowledge and ability to implement assessment good practice and recommendations stemming from research.

Table 36 below has been created in order to summarise how the results of the research undertaken will be presented in this chapter and to show how the findings will be linked effectively back to the original research questions. As the research questions seek to investigate the extent to which EAP teachers' views on EAP testing and their assessment practices reflect language testing research and EAP Assessment Literacy, questions for both the questionnaire and the interview have been generated which feed into these key areas, as identified in the Literature Review.

Table 36 Overview of research findings and linkage to research questions.

Sets of Findings in Chapter 4	Link to Research Questions
Set A1: Features of questionnaire respondent population and identification of a relevant population for the study • Details of questionnaire respondents and dropouts • Ethical consent and data protection compliance (AQ1-AQ5) • Respondent population characteristics, demographic cross tabulations and charts (AQ6-AQ12- see Figure 15 on page 155) • Identifying a relevant sample from within the total population of respondents ○ Chi Square tests to examine variables pertaining to major sub populations	Sets A1 and B1 introduce each of the main data sets from the questionnaire and survey respectively. Sets A2-A4 and B2 will then be presented in a manner which links with the research questions as listed below: *-To what extent do EAP teacher views on EAP testing and assessment practices reflect language testing research and practices which comprise Assessment Literacy?* *-How can EAP Assessment Literacy be sustained or enhanced?*
Set A2: Descriptive statistics for individual questionnaire items: • Frequency of response to individual items, Discrete Visual Analogue Scales (see sub items for AQ13, AQ14, AQ16, AQ19, AQ20, AQ22, AQ23, AQ25 and AQ26) • Rankings relating to key stages of test design (see AQ17)	
Set A3: Composite results of questionnaire Likert scales and associated statistics(see AQ13, AQ14, AQ16, AQ19, AQ20, AQ22, AQ23, AQ25 and AQ26): • Results of Factor Analysis for Likert scales • Results of Cronbach's Alpha for Likert scales • Findings from composite Likert scales	
Set A4: Content analysis of open questionnaire responses (See AQ15, AQ18, AQ21, AQ24 and AQ27): • Sample characteristics • Description of responses open questions	
Set B1: Interview data analysis- sample population description • Interviewee population characteristics and stratification in line with questionnaire population	
Set B2: Interview interactional profiles • Presentation of results stemming from interview Interactional Profiles	
Set B3: Content analysis of interview question responses BQ1-BQ18 • Description of responses to interview questions • Examples of interviewee comments	

4.2 Set A1- Features of questionnaire respondent population and identification of a relevant population for the study

4.2.1 Section introduction

In this section of the findings, results will initially be presented relating to the overall questionnaire respondent population. Features of the population of fully completed questionnaires will be considered and details associated with the number of dropouts will be displayed. Data relating to respondent ethical consent and confirmation of understanding of researcher data protection compliance will be presented. Using the data collected from AQ6-AQ12, features of the total population of fully completed questionnaires will then be revealed. Although it is recognised that the Findings chapter is distinct from the Discussion and Analysis section of the thesis, in order to present relevant results in the remaining section of the Findings, the information collected and presented in this section will then be used in order to exclude certain groups of respondent who the researcher considers to be outside the target population of the study. This will allow the identification of a more relevant sub-population which will be used for the purpose of the main findings and ensuing discussion and analysis in the next chapter.

4.2.2 Questionnaire respondents and dropouts

The total number of respondents who completed all compulsory elements of the questionnaire and who submitted all 9 sections of the survey amounted to 187. All items in the questionnaire were set as compulsory with the exception of open response items AQ15, AQ18, AQ21, AQ24 and AQ27 which invited additional responses from respondents who wished to add them at the end of each major section of the questionnaire.

As shown below in Table 37, the total number of individuals who started the questionnaire amounted to 259 indicating that those who completed and submitted all 9 sections of this research

tool constituted 72.2% (187 individuals) of those who initially started the questionnaire. Of the total number of dropouts, the available data reveals that for those 60 dropouts who completed sections as far as or beyond the gender related question, 76% (46) were female and 24% (14) were male. As will be noted in due course, the higher percentage of female dropouts also reflects the higher proportion of female respondents who completed the questionnaire.

Table 37 Cross tabulation of dropouts by section and gender

Crosstabulation of Survey Dropouts (by survey section) and Gender showing the remaining number of fully completed surveys after each survey section.				
Section in which respondents dropped out	Total number of dropouts during section	Gender		Remaining fully complete surveys related by % to total respondents
		Female	Male	
2	0	NA	NA	259 (100%)
3	12	?*	?*	247 (95.3%)
4	21	17	4	226 (87.2%)
5	27	18	9	199 (76.8%)
6	5	5	0	194 (74.9%)
7	3	3	0	191(73.7%)
8	4	3	1	187 (72.2%)
9	0	NA	NA	187 (72.2%)

*** As this section contained the gender-related item (AQ6) and as some respondents failed to complete this item, the number of dropouts in each gender group cannot be calculated.**

The researcher has taken the decision not to include partial data sets in the Findings or Discussion and Analysis Sections of this given the fact that 83%, or 60 individuals from a total of 72 dropouts withdrew from the questionnaire during the first four pages of questions. The first two pages of these questions aslo contained only personal details rather than responses from main questions related to the principal focus of study.

4.2.3 Results relating to ethical consent and data protection

As a first step in considering the findings of the overall dataset, it should be noted that items on page 2 of the questionnaire, AQ1-AQ5, were designed to request data usage permissions and

confirm respondent understanding of the researcher's compliance with data protection and ethical regulations. Of the 187 individuals who completed and submitted all compulsory items within the total survey, no concerns were raised nor were the researcher's data usage permissions withdrawn. All respondents who completed the questionnaire acknowledged having had access to the research information sheet which was available online for view and download. In addition, all respondents acknowledged voluntary participation.

4.2.4 Total population characteristics

In order to understand the features of the respondent population and data set collected via the questionnaire research tool, it is useful to examine the findings of items AQ6-AQ12 which appeared on Page 3 of the questionnaire and which were designed to collect information relating to respondent demographics. The results from this section of the questionnaire will allow an appropriate population to be identified and will reveal important respondent characteristics.

Table 38 below shows the distribution of male and female respondents (based on item AQ6) according to each of the five age brackets in item AQ7. The distribution of males and females shows that 64% of the respondents were female whilst 36% were male.

Table 38 Cross tabulation of respondent gender (AQ6) and age (AQ7)

Cross tabulation (for fully completed surveys) of respondent Age and Gender showing percentages within each gender and overall Age spread.			
Age	Gender		Total
	Female	Male	
22 to 34	15 (12.6%)	9 (13.2%)	24 (12.8%)
35 to 44	33 (27.2%)	18 (26.5%)	51(27.3%)
45 to 54	37 (31.1%)	22 (32.4%)	59 (31.6%)
55 to 64	30 (25.2%)	17 (25.0%)	47 (25.1%)
65 and Over	4 (3.4%)	2 (2.9%)	6 (3.2%)
Total	119(64%)	68(36%)	187(100%)

A notably higher number of female respondents compared to males, reflects the balance across the language teaching sector as acknowledged by Master (2005, in Hinkel, 2005, p.101-102) and Byram (2004, p.229-230). However, the results of items AQ6 and AQ7 as displayed in Table 38 above, indicate that, within each age bracket, the percentage number of respondents, based on the total number of male and female respondents, is very similar, in each case with less than one percent difference. Consequently, whether considered by each gender or as a whole group, irrespective of gender, the modal age bracket of all 187 respondents lies in the 45-54 category. As the number of respondents in the 35-44 age group is also high and similar in frequency of response to the 45-54 grouping, across both genders and the group as a whole, the population could be considered to be bimodal in nature.

In terms of respondent nationality, the results of item AQ8, as presented in Table 39 below show that respondents from 25 different nationalities completed the questionnaire with 59% of respondents (111 individuals in total) claiming to be of UK nationality and 41% of respondents (76 individuals) representing nationalities from the rest of the world.

Table 39 Cross tabulation of respondent gender (AQ6) and nationality (AQ8)

Nationality (Alphabetically)	Gender				Overall Total	Overall Total (as %)
	Female		Male			
	Total	%	Total	%		
Armenia	1	0.8	0	0.0	1	0.5
Australia	10	8.4	4	5.9	14	7.5
Canada	1	0.8	5	7.4	6	3.2
China	1	0.8	1	1.5	2	1.1
Costa Rica	1	0.8	0	0.0	1	0.5
France	2	1.7	1	1.5	3	1.6
Germany	4	3.4	0	0.0	4	2.1
Greece	1	0.8	0	0.0	1	0.5
Hungary	2	1.7	0	0.0	2	1.1
India	1	0.8	1	1.5	2	1.1
Republic of Ireland	2	1.7	2	2.9	4	2.1
Italy	1	0.8	0	0.0	1	0.5
Kazakhstan	1	0.8	0	0.0	1	0.5
Malaysian	0	0.0	1	1.5	1	0.5
Mauritius	1	0.8	0	0.0	1	0.5
New Zealand	1	0.8	2	2.9	3	1.6
Nigeria	1	0.8	0	0.0	1	0.5
Poland	1	0.8	0	0.0	1	0.5
Senegal	0	0.0	1	1.5	1	0.5
Slovenia	1	0.8	0	0.0	1	0.5
South Africa	1	0.8	1	1.5	2	1.1
Sudan	1	0.8	0	0.0	1	0.5
Taiwan	0	0.0	1	1.5	1	0.5
United Kingdom	65	54.6	46	67.6	111	59.4
United States	19	16.0	2	2.9	21	11.2
Total	119	100%	68	100%	187	100%

When a cross tabulation of gender against nationality is performed, it is revealed that the nationality with the highest reported incidence is the UK across both female and male genders with 54.6% of all females and 67.6% of all males claiming to be of UK nationality. These figures indicate that proportionally more of the non-UK respondents were of female gender.

When considering the spread of gender across other respondent nationality groups, Table 40 below indicates, in rank order and by gender the most frequently occurring nationalities of respondents. Nationalities which occur with a frequency of greater than 1 in both gender groupings are, in alphabetical order: Australia, Republic of Ireland, United Kingdom and United States. This frequency is likely to be linked to the status of these countries as major English speaking nations. Whilst New Zealand and Canadian respondents featured in the list of male respondents, with a frequency of two respondents or more, there were no female respondents from

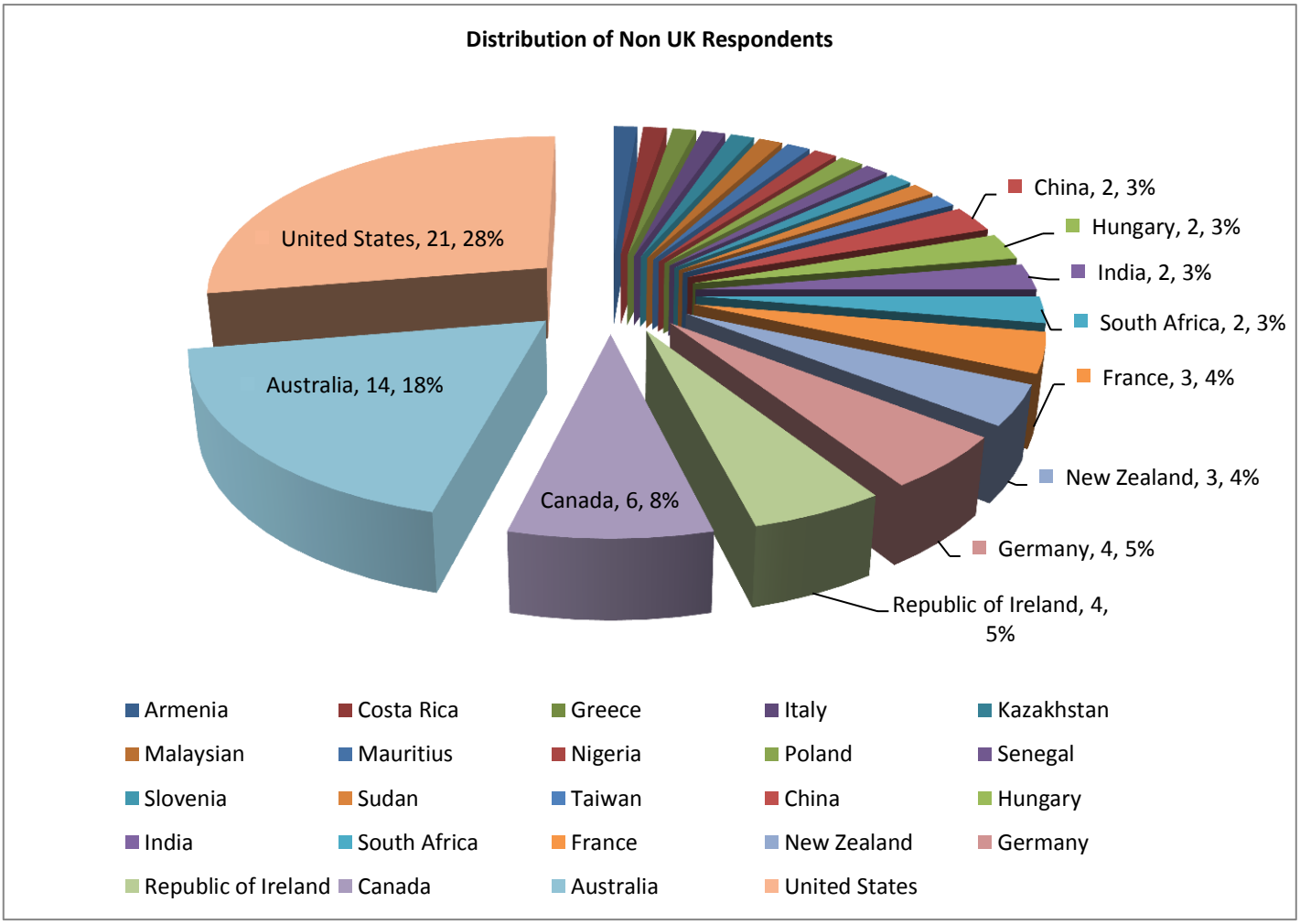
these countries with a frequency of more than one. For 15 different nationalities there are 15 lone female respondents and for 7 different nationality groups 7 lone male respondents were recorded. Other features of the data show small groupings of female respondents from Germany, France and Hungary. When comparing the two data sets, the key features which are immediately apparent are the dominance of UK respondents across both genders and the imbalance of male and female respondents from the USA. This data shows that of the total number of female respondents, those indicating US as their nationality represented 19 individuals and 16% whilst there were only two males from the US amounting to 2.9% of the total number of males.

Table 40 Ranked frequency of respondent nationality (AQ8) (for frequencies of two or more) by gender (AQ6)

Gender			
Female		Male	
Nationality	Total (% of females)	Nationality	Total (% of males)
United Kingdom	65 (54.6%)	United Kingdom	46 (67.6%)
United States	19 (16%)	Canada	5(7.4%)
Australia	10(8.4%)	Australia	4(5.9%)
Germany	4(3.4%)	Republic of Ireland	2(2.9%)
France	2(1.7%)	New Zealand	2(2.9%)
Hungary	2(1.7%)	United States	2(2.9%)
Republic of Ireland	2(1.7%)		

As respondents of UK nationality dominate respondent nationality groupings, Chart 1 below displays the distribution of Non-UK respondents. Nationality is presented irrespective of gender in order to highlight nationality groupings as a whole, with a total frequency of two or more.

Chart 1 Distribution of non UK respondents by nationality



The data shows that 11 nationalities are represented with a frequency of two or more respondents. 'Other nationalities' collectively describes 13 respondents who are lone representatives of a nationality.

Table 41 Spread of respondent nationalities (AQ8) from outside the UK

Ranked frequency of non UK Nationalities	Total (% of non UK respondent nationalities)
(1)United States	21(27.6%)
(2) Australia	14(18.4%)
(4) Other Nationalities*	13(17.1%)
(3) Canada	6(7.9%)
(5) Germany	4(5.3%)
(6) Republic of Ireland	4(5.3%)
(7) France	3(3.9%)
(8) New Zealand	3(3.9%)
(9) China	2(2.6%)
(10) Hungary	2(2.6%)
(11) India	2(2.6%)
(12) South Africa	2(2.6%)

***indicates all other nationalities with only one single respondent**

The spread of respondent nationalities from outside the UK is summarised in Table 41 above which indicates that the three most frequently cited nationalities of respondent outside the UK are US, Australia and Canada which collectively account for more than 40% of all non-UK nationality respondents. Once again, this prevalence is arguably unsurprising, given the means of disseminating the questionnaire amongst professional organisations associated with EAP which have global membership but notably occupy key bases in the North America and Australia. These nationalities are also based within the 'inner circle' and represent some of the traditional bases of English where it is a primary language (Crystal, 2003, p.60-61) and where EAP is also taught most widely. Markedly, the Republic of Ireland, New Zealand and South Africa which also are listed in Table 41 fall into this same 'inner circle' category. Given the international nature of English language teaching which includes practices associated with EAP and its assessment, respondent nationality is a somewhat imprecise means of categorising the population of respondents associated with this research project, consequently item AQ10 was created in order to

investigate respondents' most recent country of EAP practice. Chart 2 and Table 42 present this data below.

Chart 2 Distribution of countries where respondents have been most recently involved with EAP testing and assessment

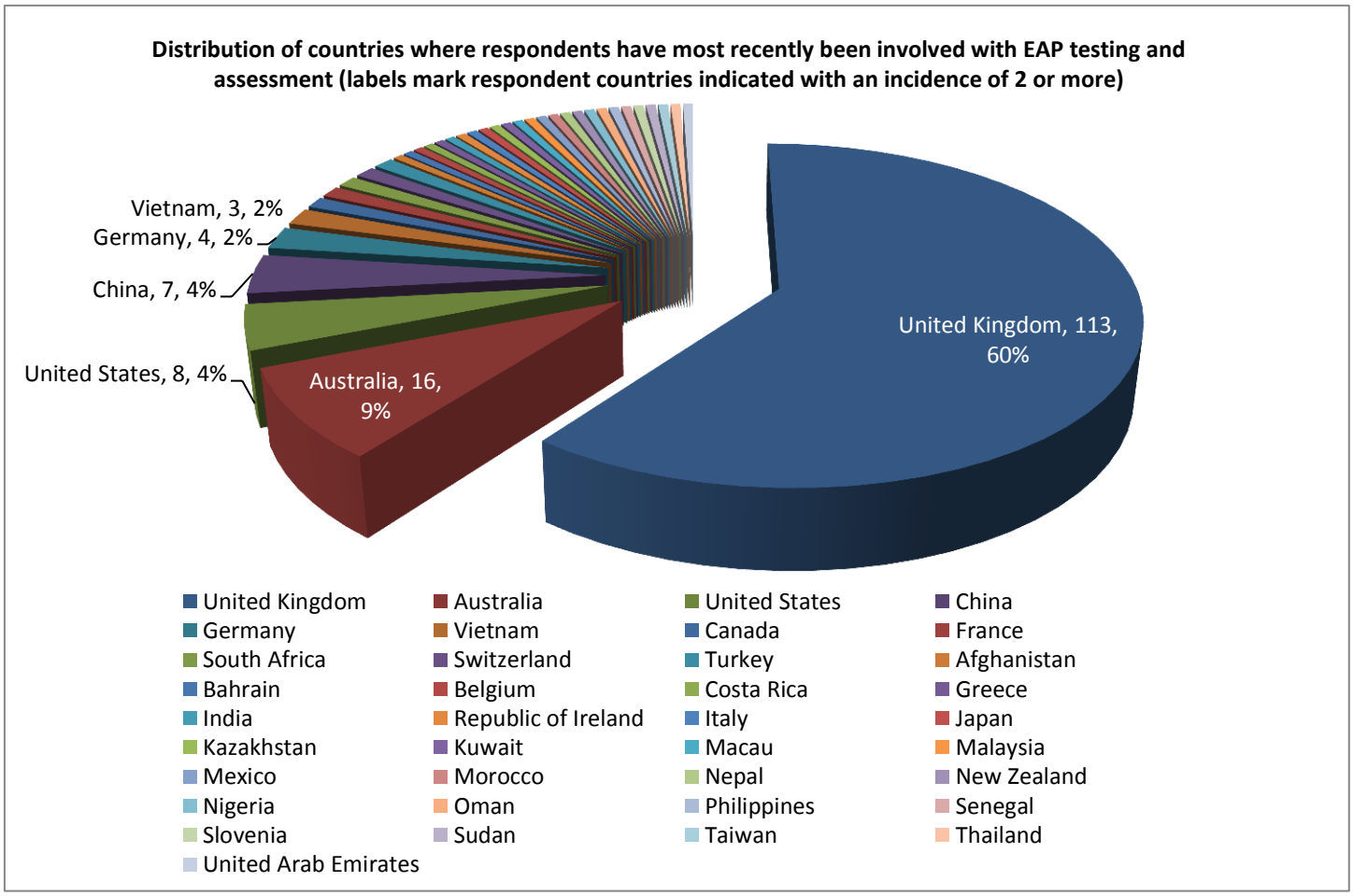


Table 42 Cross tabulation of respondent gender (AQ6) and country of most recent involvement with EAP testing and assessment (AQ10)

Country of most recent involvement with EAP testing and assessment (Alphabetically)	Gender				Overall Total	Overall Total (as %)
	Female		Male			
	Total	%	Total	%		
Afghanistan	1	0.8	0	0.0	1	0.5
Australia	15	12.6	1	1.5	16	8.6
Bahrain	1	0.8	0	0.0	1	0.5
Belgium	1	0.8	0	0.0	1	0.5
Canada	1	0.8	1	1.5	2	1.1
China	4	3.4	3	4.4	7	3.7
Costa Rica	1	0.8	0	0.0	1	0.5
France	1	0.8	1	1.5	2	1.1
Germany	4	3.4	0	0.0	4	2.1
Greece	1	0.8	0	0.0	1	0.5
India	0	0.0	1	1.5	1	0.5
Republic of Ireland	1	0.8	0	0.0	1	0.5
Italy	1	0.8	0	0.0	1	0.5
Japan	0	0.0	1	1.5	1	0.5
Kazakhstan	1	0.8	0	0.0	1	0.5
Kuwait	0	0.0	1	1.5	1	0.5
Macau	0	0.0	1	1.5	1	0.5
Malaysia	0	0.0	1	1.5	1	0.5
Mexico	0	0.0	1	1.5	1	0.5
Morocco	1	0.8	0	0.0	1	0.5
Nepal	0	0.0	1	1.5	1	0.5
New Zealand	0	0.0	1	1.5	1	0.5
Nigeria	1	0.8	0	0.0	1	0.5
Oman	0	0.0	1	1.5	1	0.5
Philippines	1	0.8	0	0.0	1	0.5
Senegal	0	0.0	1	1.5	1	0.5
Slovenia	1	0.8	0	0.0	1	0.5
South Africa	0	0.0	2	2.9	2	1.1
Sudan	1	0.8	0	0.0	1	0.5
Switzerland	2	1.7	0	0.0	2	1.1
Taiwan	0	0.0	1	1.5	1	0.5
Thailand	0	0.0	1	1.5	1	0.5
Turkey	1	0.8	1	1.5	2	1.1
United Arab Emirates	0	0.0	1	1.5	1	0.5
United Kingdom	68	57.1	45	66.2	113	60.4
United States	8	6.7	0	0.0	8	4.3
Vietnam	2	1.7	1	1.5	3	1.6
Total	119	100	68	100	187	100

When gender and country of most recent involvement with EAP assessment is cross tabulated (see Table 43 below) it is clear that the country with the highest reported incidence is the UK across both female and male genders with 57.1% of all females and 66.2% of all males explaining that they have most recently been involved with EAP testing and assessment in the UK. These figures show that that a higher proportion of male respondents (42.9%) have gained their most recent experience of EAP assessment outside the UK, when compared to the proportion of female respondents (37.8%). When considering the spread of gender across other respondent countries of most recent EAP assessment practice, Table 43 below summarises, in rank order, countries of most recent EAP assessment practice which occur with a frequency of greater than 1. In both gender groupings the only two countries which appear in association with both genders are UK and China. Other features of the data in Table 43 show that proportionally a larger percentage of the male respondents (23.9% as opposed to 2.5% for females) have gained their EAP assessment experience in a country for which they were the lone male respondent.

Table 43 Ranked frequency of respondent most recent country of practice in EAP assessment (AQ10) (for frequencies of two or more) by gender (AQ6)

Gender			
Female		Male	
Nationality	Total (% of females)	Nationality	Total (% of males)
United Kingdom	68 (57.1%)	United Kingdom	45 (66.2%)
Australia	15 (12.6%)	China	3 (7.0%)
United States	8(6.7%)	South Africa	2 (2.9%)
China	4(3.4%)		
Germany	4(3.4%)		
Vietnam	2(1.7%)		
Switzerland	2(1.7%)		

When country of most recent involvement with EAP testing and assessment is cross tabulated against country of nationality (NB this table has not been included due to its dimensions) the most marked frequency of occurrence is UK nationals who have most recently been involved with EAP

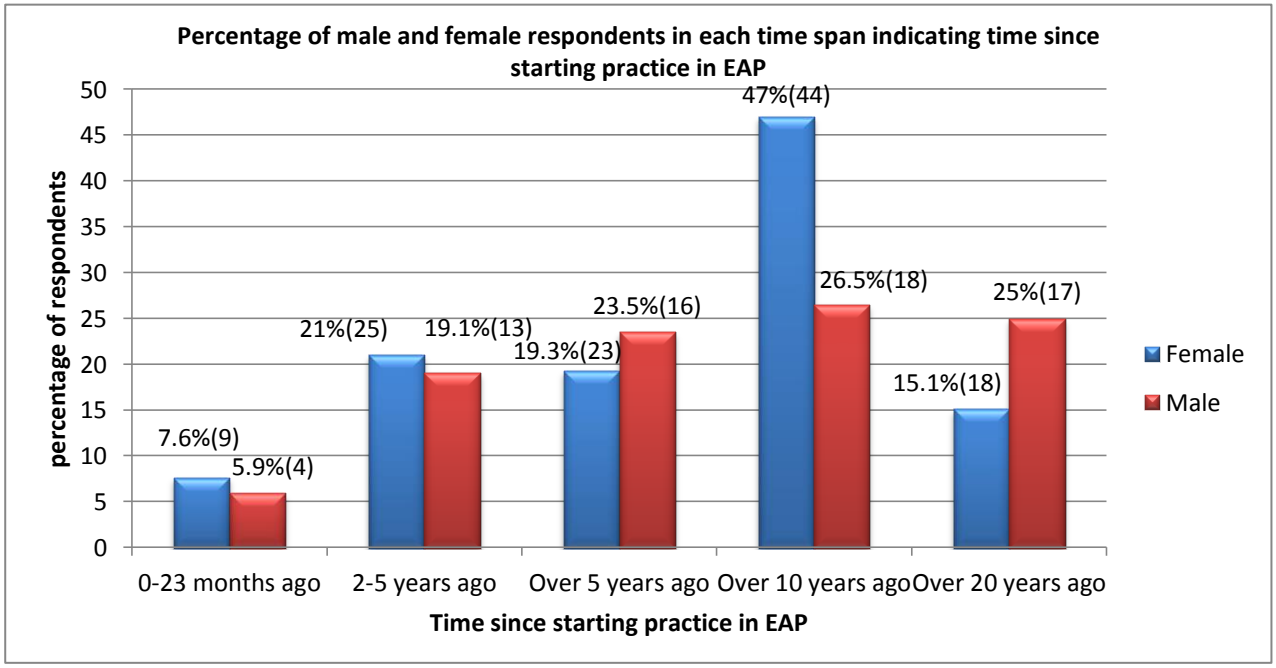
assessment in the UK, with a total of 92 respondents in this situation. The next highest clustering of nationality shows 10 Australian nationals who have most recently been associated with EAP testing and assessment in Australia and 7 US nationals who have most recently been associated with EAP testing and assessment in the US. Clusters of nationalities based outside their home countries show 6 US nationals who have been most recently associated with the field in the UK. 5 UK nationals are most recently associated with EAP testing and assessment in China. Similarly, 3 Australian and 3 Republic of Ireland nationals have most recently been associated with the field in the UK.

Turning to length of service in the field of EAP, Table 44 and Chart 3 below show cross tabulations of respondents' most recent institution of practice against both gender and length of service in EAP.

Table 44 Cross tabulation of gender and time since starting practice in EAP

Time since starting practice in EA	Gender		
	Female (%)	Male (%)	Total (%)
0-23 months ago	9 (7.6%)	4(5.9%)	13 (7.0%)
2-5 years ago	25(21.0%)	13(19.1%)	38(20.3%)
Over 5 years ago	23(19.3%)	16(23.5%)	39(20.9%)
Over 10 years ago	44(47.0%)	18(26.5%)	62(33.2%)
Over 20 years ago	18(15.1%)	17(25.0%)	35(18.7%)
Total	119(100%)	68(100%)	187(100%)

Chart 3 Males and female respondents according to time since starting practice in EAP



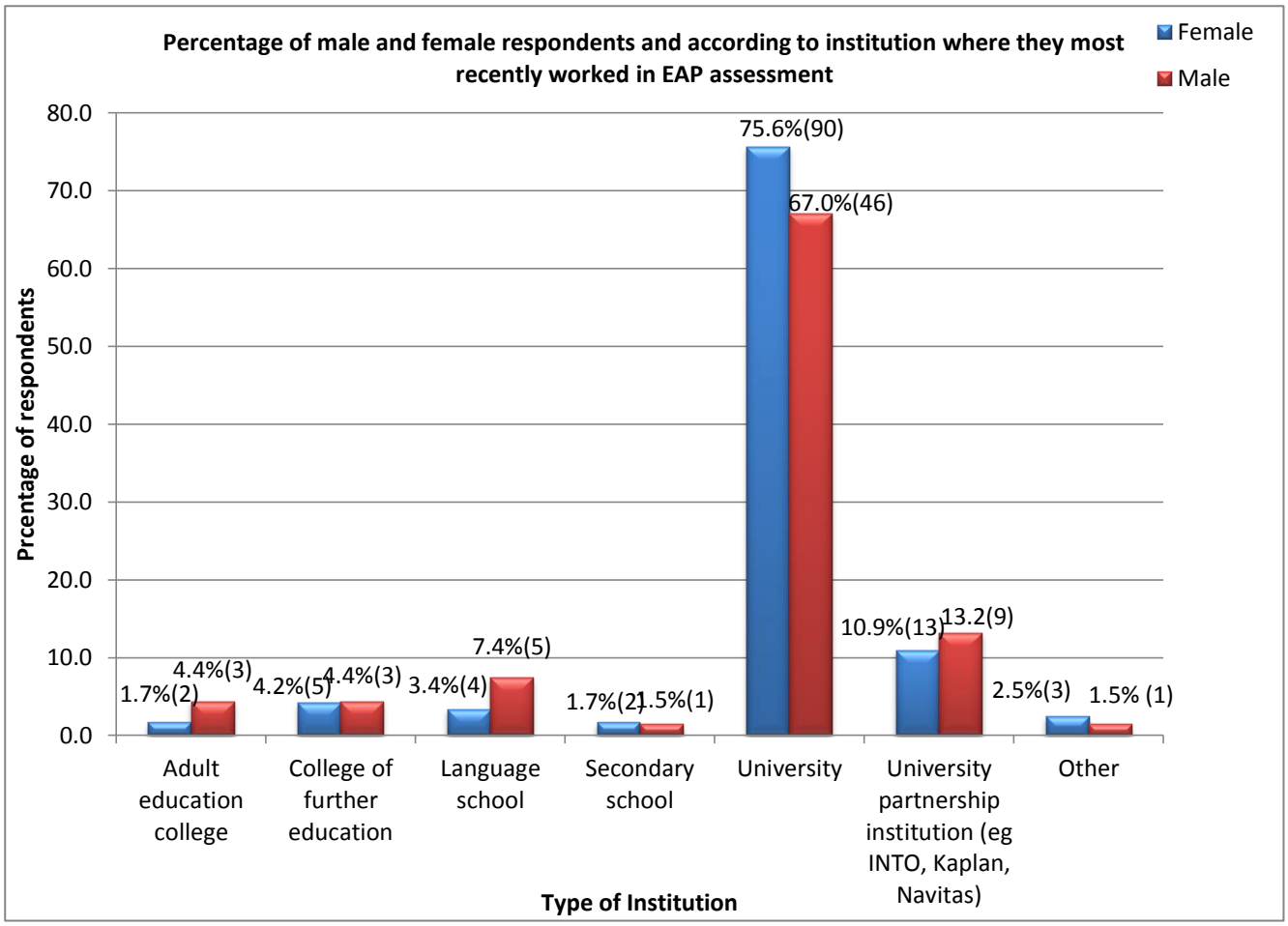
As the data suggests, the proportion of male and female respondents in each of the three time bands, from ‘0-23 months’ to ‘over 5 years’, shows a similar pattern for both genders with the percentage of males and females in each grouping representing broadly similar proportions of the overall group. The most marked observation can be observed in the ‘Over 10 years ‘ and ‘over 20 years’ categories where the proportion of female respondents peaks at 47% in the ‘over 10 years’ category, which represents the clear modal response and then falls back sharply during the ‘over 20 years’ band. In contrast, the spread for male respondents demonstrates a pattern which can better be described as multimodal across the three categories from ‘over 5 years to ‘over 20 years’ which range from 23.5% to 26%. Notably in the ‘over 20 years’ category the actual number of male and female respondents is very similar at 17 and 18 respectively, despite the overall majority of female respondents at 119 compared to 68 males.

When considering institutions were respondents have most recently worked in EAP assessment, Table 45 and Chart 4 show cross tabulations according to institutions where respondents have most recently worked in EAP assessment, in addition to the related gender division.

Table 45 Cross tabulation of respondent gender and institution most recently involved in EAP testing and assessment

Gender	Most recent institution where involved with EAP testing and assessment							Total
	Adult education college	College of further education	Language school	Secondary school	University	University partnership institution (eg INTO, Kaplan, Navitas)	Other	
Female	2 (1.7%)	5 (4.2%)	4 (3.4%)	2 (1.7%)	90 (75.6%)	13 (10.9%)	3 (2.5%)	119(100%)
Male	3 (4.4%)	3 (4.4%)	5 (7.4%)	1 (1.5%)	46 (67.6%)	9 (13.2%)	1 (1.5%)	68 (100%)
Overall Total	5 (2.7%)	8 (4.3%)	9 (4.8%)	3 (1.6%)	136 (72%)	22 (11.7%)	4 (2.1%)	187 (100%)

Chart 4 Male and female respondents according to institution where they have most recently worked in EAP assessment



As this data demonstrates, the type of institution which is most frequently cited by males and females as the institution where they have most recently been involved with EAP assessment is

‘University’ with figures of 75.6% and 67% for females and males respectively. This information shows that a greater proportion of male respondents, compared to female counterparts have worked in EAP assessments outside of institutions described as universities. The next most frequent response for both male and female respondents was ‘University partnership institution’ . When considered collectively this returns figures of 13.5% of females and 29.8% of males who work outside of the ‘University’ or ‘University Partnership’ sectors.

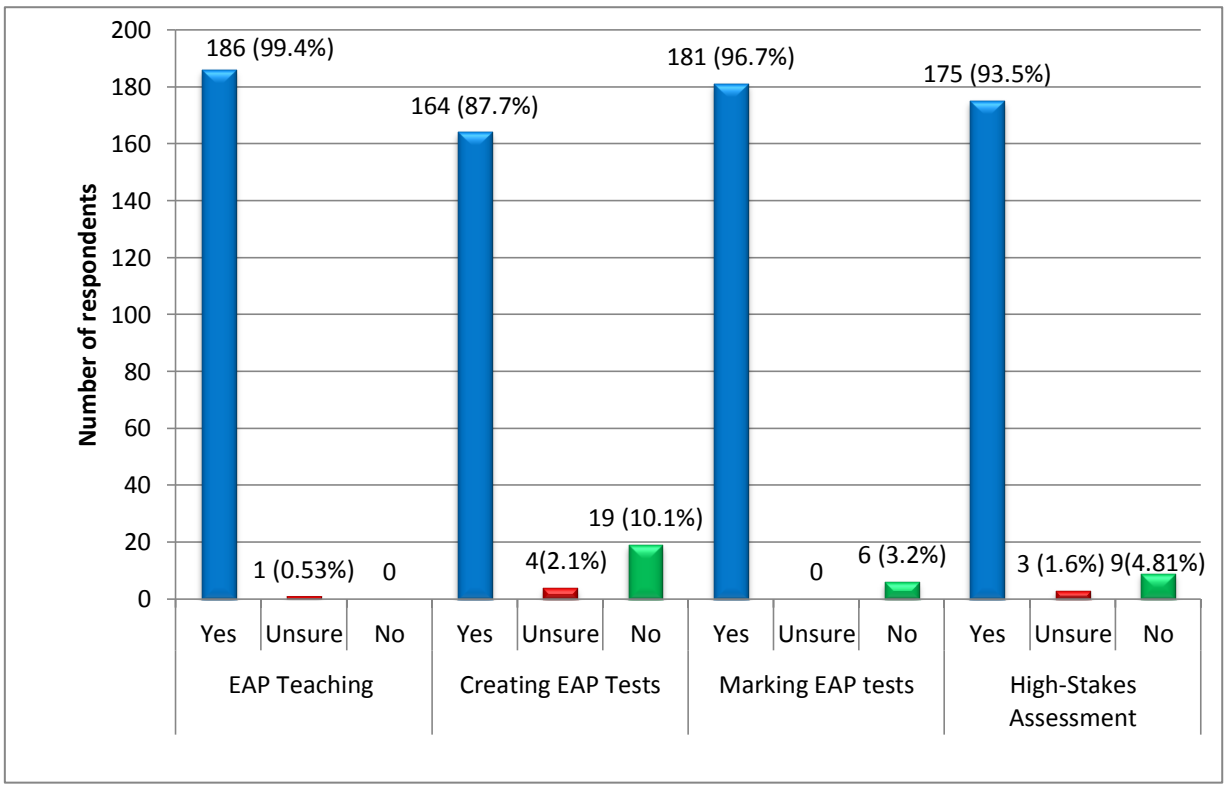
Table 46 below shows to the amount of time that respondents have spent working in the field of EAP against the type of institution where respondents have most recently worked in EAP assessment. The data shows that in each age bracket the modal response falls in the ‘University’ institutional category. For each age bracket the second most frequent response was ‘University partnership institution.’

Table 46 Cross tabulation of respondent time since starting work in EAP and type of institution where most recently involved with EAP assessment

Crosstabulation of respondent time since starting work in EAP and institution most recently involved in EAP testing and assessment								
Time since starting work in EAP	Type of institution where most recently involved in EAP testing and assessment							Totals
	Secondary School	College of further education	Adult education college	Language School	University	University partnership institution (e.g. INTO, Kaplan, Navitas)	Other	
0-23 months	0	0	2	0	8	3	0	13
2-5 years	2	2	2	1	26	5	0	38
Over 5 years	0	3	0	4	27	5	0	39
Over 10 years	0	1	1	1	51	5	3	62
Over 20 years	1	2	0	3	24	4	1	35
Totals	3	8	5	9	136	22	4	187

Item AQ9, shown in Chart 5 below, collected information relating to respondent experience of EAP and EAP assessment, in order to determine if respondents have relevant familiarity with the field. As indicated, in each area of experience, a high proportion of the total number of respondents confirmed their involvement in practices associated with EAP teaching and assessment. In addition, when the total number of respondents answering 'Yes' to all four areas of experience is considered, it can be seen that 82% of all respondents (155 individuals) chose to confirm their experience in each of the respective areas. Results of respondents answering 'No' to two or more of the four areas of experience identified only two respondents with multiple areas of inexperience and no respondents answering 'no' to all four sub-items in AQ9. As a result, no respondents identified a comprehensive lack of experience by answering 'No' to each of the four listed areas of EAP and EAP assessment experience.

Chart 5 Respondent experience of EAP and EAP assessment



4.2.5 Filtration of the total number of respondents to identify a suitable and representative sample

The findings presented in Section 4.2 present the features of the overall data set for the 187 respondents who completed all compulsory sections of the questionnaire. For the subsequent stages of the presentation of findings and analysis, and in order to focus on a sample or population which can be considered adequately homogenous, certain decisions have been made regarding which categories of respondent to include for further consideration. This represents part of the process of reducing variables in the dataset, so that key features can be identified. It also helps to demonstrate that the sample used is sufficiently similar to the identified target population (Dörnyei, 2003, p.107-108). This aspect of the research is recognised by the researcher as crucial, given that it is necessary for the sample to be representative of a wider population in order for subsequent conclusions to be meaningful and generalisable (Dörnyei, 2007, p.96).

Figure 23, below on page 209, shows the variables which were given key consideration when deciding which groups of respondents to filter out of the main findings and analysis.

The first step involved identifying whether any respondents had withheld their consent for participation in the research or use of their data in items AQ1-AQ5. As no respondents withheld their consent, this was not a factor which required further consideration. The variables associated with respondent gender, age and nationality in AQ6-AQ8 were not used in the process of filtering the data set in order to identify a reduced population, as it was felt that this could have led to an ethically inappropriate selection process.

The second stage which was considered for filtration purposes involved responses to AQ9 where respondents were asked to comment on four aspects of their experience of EAP practice and assessment. As a very high percentage of respondents indicated experience in each of the four

listed areas and no respondents identified a complete lack of experience, this information was not used to screen the data set or exclude any respondents from the main population.

The results of AQ10 showed a division of respondents based on their involvement with EAP in the UK and The rest of the world. Although the grouping identified as the rest of the world demonstrated further clusters of respondent, the researcher has decided to consider it as a single category given the large number of respondents who are lone representatives of a single nationality. Although it was also considered whether the main population should exclude international practitioners completely, given the large number of respondents based in the UK, it was judged as relevant and important that respondents based in international locations should be included in the study, given the international and globally mobile nature of contemporary EAP (Alexander et al., 2008, p.5).

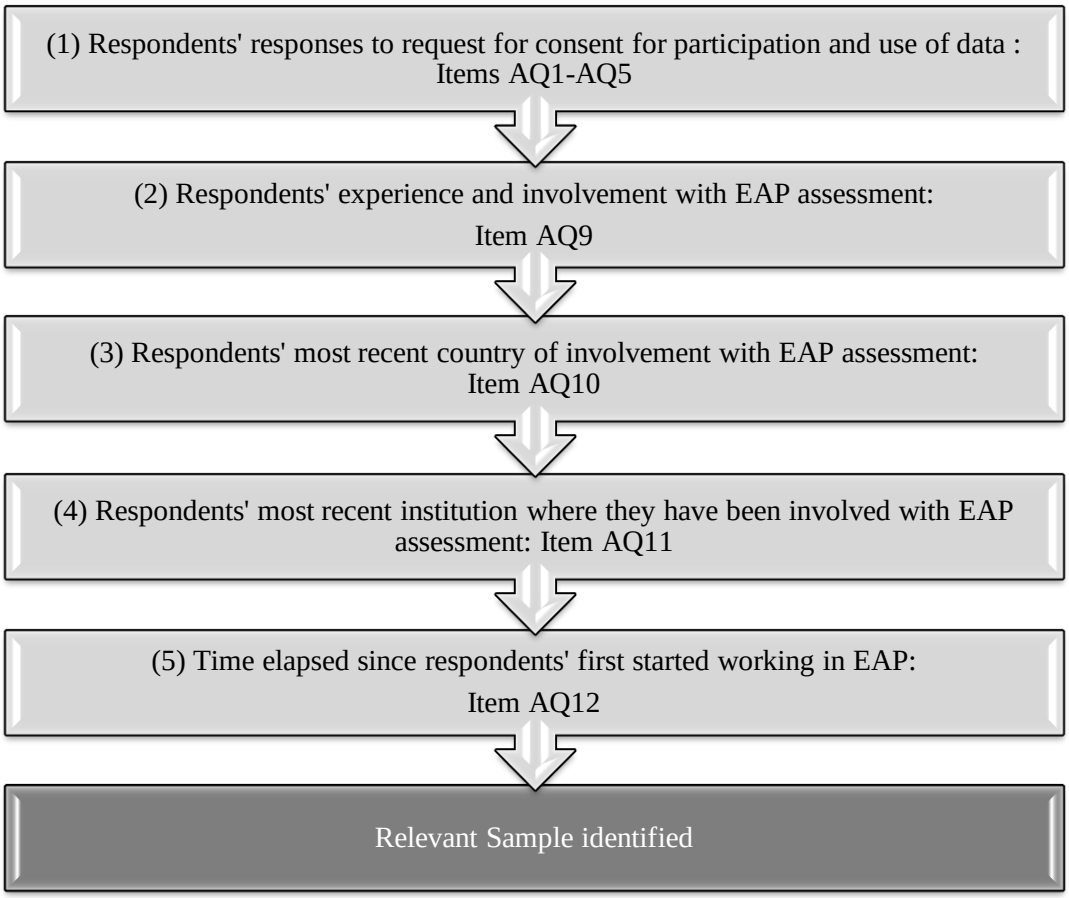
With regard to respondents' most recent institutional involvement with EAP assessment, the results of AQ11 showed that a high proportion of respondents have most recently been involved with EAP assessment in either a university or university partnership institution. Given this clustering and the researcher's concern to base the study on respondents who have experience of working in an EAP context which is closely linked to University study, it was decided that respondents who have most recently been involved in EAP at institutions other than universities or university partnerships should be excluded from the population used for the main findings of this research.

It was also considered whether the data from AQ12 should be used in order to exclude respondents who have the least amount of experience in EAP, however finally it was decided that the main population should not be filtered according to years of experience as this would

unsatisfactorily represent the true features of a representative group of practising EAP teachers and would thus restrict generalisation.

Based on the decisions above which have led to the filtration of the main data set, comprised of 187 respondents, a remaining total of 158 respondents has been identified which represents respondents who have most recently been involved with EAP assessment either in universities or university partnership institutions, either in the UK or across the rest of the world.

Figure 23 Considerations for filtration of the full respondent population



In order to confirm whether the data sample, based on respondents who have most recently worked in EAP at either universities or university partnership organisations either in the UK or overseas, can be considered as a single population, the similarity or differences pertaining to these

variables warrants additional scrutiny. Unfortunately, the differences displayed in a table alone do not enable the researcher to determine whether fluctuations occur by chance or whether there are any interesting or noteworthy patterns. Consequently frequency comparison statistics can be used in order to determine the probability that frequencies depart from what would be expected in a given table (Brown, 2001, p.158). One statistic which is useful for this purpose and which is suitable for use with ordinal/categorical data is Chi-Square (χ^2). In order for Chi-Square to be undertaken certain provisos were confirmed (Brown, 2001, p.168-169) to ensure that the data was suitable for use with this statistic. These provisos include:

(1) Use with nominal data

(2) Use only with variables which are independent from each other

(3) Expected frequencies are greater than 5

As the researcher has judged that the data in question, drawn from items AQ6-AQ12, conform to the requirements 1 and 2 above, it was decided to proceed with Chi-Square statistics and to consider how to deal with the situation should expected frequencies as revealed by the software output require additional attention.

Chi-Square was then calculated for a set of seven pairs (C-S1-CS2) of cross tabulated variables as described below:

(C-S1) Respondent gender and country of most recent involvement with EAP assessment

(C-S2) Respondent age group and country of most recent involvement with EAP assessment

(C-S3) Respondent institution of most recent involvement with EAP assessment and country of most recent involvement with EAP assessment

(C-S4) Respondent time elapsed since starting EAP practice and country of most recent involvement with EAP assessment

(C-S5) Respondent gender and type of institution of most recent involvement with EAP assessment

(C-S6) Respondent age group and type of institution of most recent involvement with EAP assessment

(C-S7) Respondent time elapsed since starting EAP practice and type of institution of most recent involvement with EAP assessment

Each of the frequency comparison statistics computed through the Chi-Square was predicated on a null hypothesis that the two variables in each pairing were independent of each other. In each case, the cut-off point for significance in the Pearson Chi-Square Value that has been used is 0.05 or 5%. This means that if the output tables from SPSS show that Pearson Chi-Square value is insignificant, it will show a figure of more than 0.05 and consequently the two variables can be considered to be independent rather than related. If this situation arises, the null hypothesis can be retained.

Table 47-Table 49 below relate to C-S1 and act as a sample of the data from each of the seven full sets of information gathered through Chi-Square calculations in SPSS. The full set of data for each of the Chi-Square calculations referred to in this section can be found in the Appendix of this document in section 8.6 on page 442.

A 2x2 Chi-Square analysis revealed that there was no significant relationship between respondent gender and the country of respondents' most recent involvement with EAP assessment. This is demonstrated through the χ^2 value of 3.254^a and the probability value, or p-value of 0.071, as displayed in Table 49.

Table 47 Case processing summary for (C-S1)

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Gender and the country of respondents' most recent involvement with EAP testing	158	100.0%	0	.0%	158	100.0%

Table 48 Crosstabulation of respondents' gender and the country of respondents' most recent involvement with EAP testing (C-S1)

			Respondents' country of most recent involvement with EAP testing and assessment		
			Rest of The World	United Kingdom	Total
	Female	Count	41	62	103
		Expected Count	35.9	67.1	103.0
	Male	Count	14	41	55
		Expected Count	19.1	35.9	55.0
Total		Count	55	103	158
		Expected Count	55.0	103.0	158.0

Table 49 Chi-Square Tests (C-S1)

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3.254[a]	1	.071		
Continuity Correction[b]	2.652	1	.103		
Likelihood Ratio	3.342	1	.068		
Fisher's Exact Test				.081	.050
N of Valid Cases	158				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 19.15.

b. Computed only for a 2x2 table

Table 50 below has been created in order to summarise the findings of each of the additional Chi-Square tests, C-S1-C-S7, which have been conducted and which can be viewed in full in the Appendix:

Table 50 Summary of findings from Chi-Square tests C-S1-C-S7

Codes for pairs of cross tabulated variables	Description of variables compared	Number of nominal variables compared	χ^2 value	*p*-value
C-S1	Respondent gender and country of most recent involvement with EAP assessment.	2X2	3.254[a]	.071
C-S2	Respondent age group and country of most recent involvement with EAP assessment.	4X2	2.287[a]	.515
C-S3	Respondent institution of most recent involvement with EAP assessment and country of most recent involvement with EAP assessment	2X2	3.114a	.078
C-S4	Respondent time elapsed since starting EAP practice and country of most recent involvement with EAP assessment.	4X2	2.585[a]	.460
C-S5	Respondent gender and type of institution of most recent involvement with EAP assessment.	2X2	.419[a]	.517
C-S6	Respondent age group and type of institution of most recent involvement with EAP assessment.	3X2	6.962[a]	.031
C-S7	Respondent time elapsed since starting EAP practice and type of institution of most recent involvement with EAP assessment.	2x2	1.253[a]	.263

When undertaking the Chi-Square tests, variables in C-S2, C-S4, C-S6 and C-S7 contained expected count values with less than 5 counts which required the recoding of data in order to create alternative categories with expected count values higher than this figure. Although in a number of cases expected count values differed slightly from the true count value, this difference was not considered substantial. Furthermore, *p*-values for each of C-S1, C-S2, C-S3, C-S4, C-S5, and C-S7 all reached levels of more than 0.05, showing that there is no significant relationship between the selected pairs of variables which were cross tabulated. Only C-S6 with a *p*-value of 0.031 suggests that a relationship does exist with the variables which have been cross tabulated in this test.

Chi-Square analysis revealed that there is a significant relationship between respondent age group and the type of institution where respondents have most recently been involved with EAP assessment. Most notably the relationship indicates that a much higher proportion than expected of respondents working in partnership institutions are found in the lower age bracket of 22 to 44 and fewer than expected can be found in both the higher age brackets of 45-54 and '55 and over'

When considering the implications of these findings for the handling and presentation of data in the remaining elements of this findings section, the researcher has decided to consider the population of 158 respondents who have most recently been involved with EAP assessment in either a University or University partnership institution in location categories marked as either UK or the rest of the world, to be a single and representative sample given the relative absence of significantly related variables. A more qualitative rationale which supports this sample grouping is the fact that many university partnership organisations are former university departments which have been taken over by private providers, thus including staff with the same or similar professional experience as those who work or still work for university language centres. It is postulated by the researcher that the higher number of EAP teachers in the younger age brackets, as shown by C-S6, may be representative of the employment terms and conditions pertaining to university partnership organisations, which now differ to those of universities (UCU, 2012) and which may be less appealing to more experienced or older EAP staff. Furthermore, the inclusion of practitioners based overseas at relevant institutions also reflects the geographic mobility which is inherent to the EAP profession. This population comprising the elements above will henceforth be referred to as Population B, with Population A being the 187 respondents who completed and submitted all sections of the survey.

4.2.6 Section conclusion

This part of the findings chapter has revealed key information relating to the questionnaire respondent population, and has also presented data relating to those who failed to complete the full quesationnaire tool. With regard to ethicality, evidence of participant consent has been confirmed along with the researcher's commitment to follow data protection protocols. A process of filtration in order to reduce the total population, Population A, of repsondents to a group who are more representative of the population of EAP tutors required by the research questions in this theis, Population B, has also been established. Features of the sample population will be explored further in Section 5.2 when the characteristics of filtered sample, Population B, are analysed and discussed.

4.3 Set A2-Descriptive statistics for individual questionnaire items:

4.3.1 Section introduction

As features of the overall population of respondents have now been discussed and a more relevant subpopulation of respondents has been identified in the form of Population B, this section of the findings will present descriptive statistics associated with sets of individual items which take the form of Discrete Visual Analogue Scales (DVAS), asking respondents to select an option from 1-5. This includes the results of series of items in sections AQ13, AQ14, AQ16, AQ19, AQ20, AQ22, AQ23, AQ25 and AQ26. The results of AQ17, which ranks key stages of test design, will also be presented. This section therefore reveals the respondent population's responses to individual questions which have been devised, in line with the research questions. The aim has been to identify EAP tutor views on EAP testing and assessment practices. This is undertaken with a view to determining whether the opinions expressed reflect language testing research and practices which comprise Assessment Literacy. It should be noted that Set A3 will present composite results of Likert scales in section 4.4.

4.3.2 Frequency of response to individual DVAS items AQ13, AQ14, AQ16, AQ19, AQ20, AQ22, AQ23, AQ25 and AQ26

This section will present the results for Population B relating to sets of items included in AQ13, AQ14, AQ16, AQ19, AQ20, AQ22, AQ23, AQ25 and AQ26. Figure 15 on page 155 explains the aim and structure of each set of items and the full questionnaire is included in the Appendix in 8.1 on page 420. In order to contextualise the findings referred to in this section, reference can again be made to Table 30 on page 158 which refers to key areas of EAP Assessment Literacy linked to items/sections within the research questionnaire. These areas of EAP Assesment Literacy were established during the Literature Review and summarised in Section 2.7 and can be summarised as follows:

- training, skills, strengths and weaknesses in EAP testing and assessment
- approaches to EAP testing and assessment design
- views on/ experience of validity and reliability in EAP testing and assessment
- practices involving analysis and interpretation of the results and scores of EAP tests and assessments
- consideration of ethics in EAP testing and assessment
- influence of research and other resources for the purpose of EAP testing and assessment on EAP teacher assessment practices

In order to show a linkage with the research questions, the findings in this section will be grouped according to the aspects of Assessment Literacy which they repreresent.

After each reference to percentages, the total number of respondents represented by the percentage figure is added in brackets, for further reference. In the questions and associated charts, placed

below each description of findings, a number of different scales have been used, a key for which is provided below:

Table 51 Scales used in DVAS items in Questionnaire

Scale 1	Scale 2	Scale 3	Scale 4
– Strongly Agree (SA) – Agree (A) – Neither Agree Nor Disagree (NAND) – Disagree (D) – Strongly Disagree (SD)	– Very Highly Skilled (VHS) – Highly skilled (HS) – Average skilled (AS) – Low skilled (LS) – Very low skilled (VLS)	– Very Familiar (VF) – Familiar (F) – Neither Familiar Nor Unfamiliar (NFNU) – Unfamiliar (U) – Very Unfamiliar (VU)	– Yes – Unsure – No

4.3.2.1 Responses to items linked to training, skills, strengths and weaknesses in EAP testing and assessment

4.3.2.1.1 Responses to items AQ13-A-J

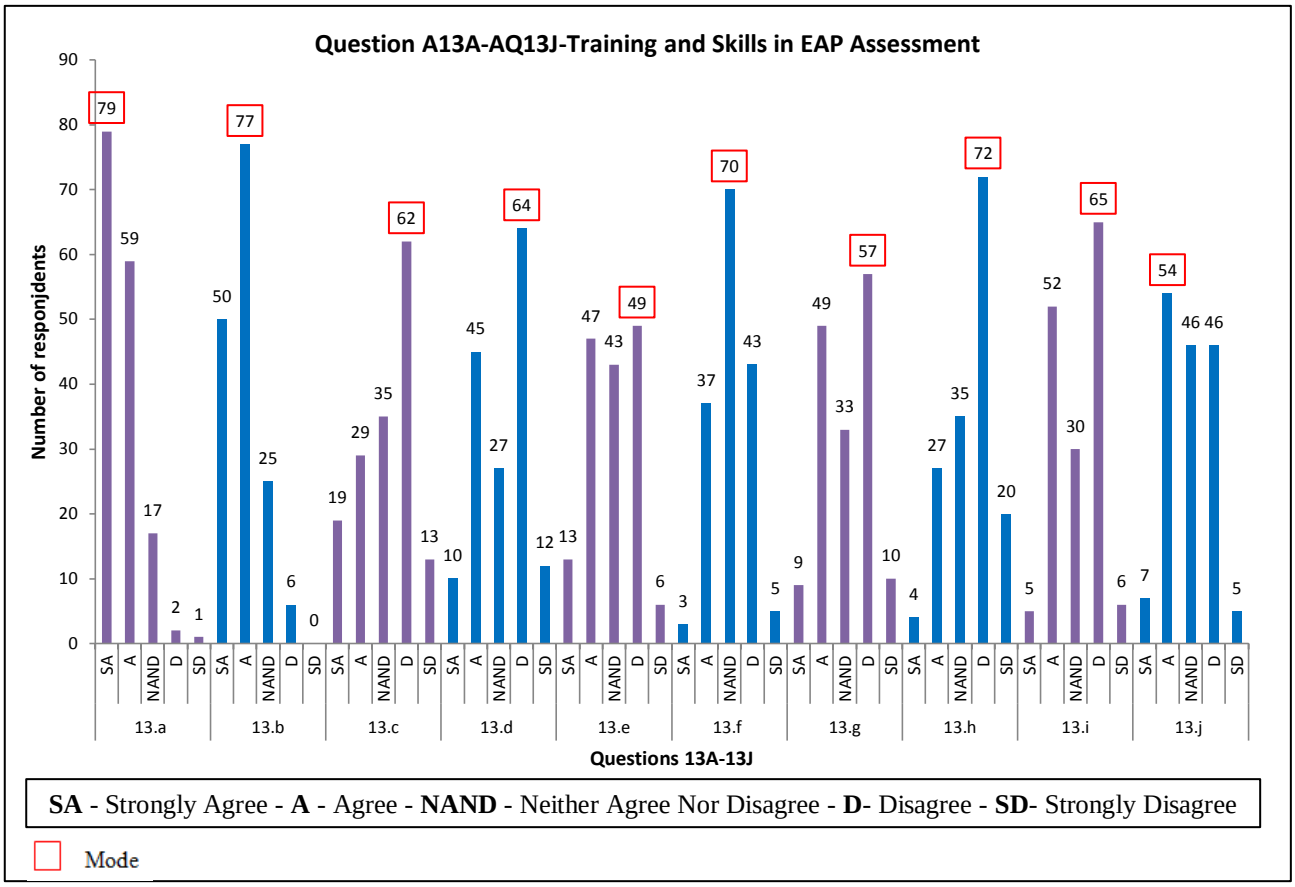
Item AQ13 included 10 sub items A-J which were designed to ask questions relating to respondent training and skills development in EAP assessment. This section aimed to identify the source of respondent assessment capabilities as well as indicators of respondent interest in the development and application of such skills in a manner which relates to the inherent proactive requirement of Assessment Literacy as defined by Stiggins (1995) and Popham (2001).Table 52 below displays the questions which were asked.

Table 52 Items/Questions in set AQ13

Item code	Item
AQ13a	My interest in EAP is driven by aspects of the field other than testing or assessment.
AQ13b	Testing and assessment skills are crucial for my role as an EAP tutor.
AQ13c	The skills that I have acquired in EAP assessment have mainly been developed through courses I have taken which include focus on language testing.
AQ13d	I regularly read books or articles on language testing.
AQ13e	EAP testing and assessment is one of my key skill areas.
AQ13f	The reference and guidance material available to me on the subject of language testing and assessment is very user-friendly.
AQ13g	I feel that I have had sufficient training in EAP assessment and testing practices.
AQ13h	It is more straightforward to test skills in EAP than to test subject knowledge.
AQ13i	It is straightforward to identify the key EAP language areas and skills that students need to be assessed in.
AQ13j	It is straightforward to recognise a question which is not working properly in an EAP test or assessment.

Chart 6 below presents the frequency of responses across items AQ13A-J.

Chart 6 Questions AQ13A-AQ13J for population B



Item AQ13a attempts to measure the level of specific interest in EAP assessment pertaining to respondents, which is arguably a key factor in realising the dynamic potential of Assessment Literacy as described by educationalists such as Price et al. (2012. p.9). The frequency of

responses shows that 87.3% (138) of respondents either agree or strongly agree that their interest in EAP is driven by aspects of the field other than testing or assessment. However, in response to AQ13b 80.3% of respondents either agree or strongly agree that skills in testing and assessment are crucial to their role in EAP.

In response to item AQ13c only 30.3% of respondents strongly agree or agree that their skills in EAP assessment and testing have been developed through courses focusing on language testing. Item AQ13d will be discussed in Section 4.3.2.5 as it judged that this item relates more closely to the influence of research on EAP teacher assessment practices.

Item AQ13e asked respondents to gauge their level of agreement with the statement that EAP testing and assessment represents one of their key skill areas, in order to gauge EAP teachers' confidence in their own abilities. Responses to this item revealed a multimodal frequency of response with 29.7% (47) agreeing, 27.2% (43) neither agreeing nor disagreeing and 31% (49) disagreeing. Notably only 37.9% (60) respondents agreed or strongly agreed that EAP testing and assessment represented a key skill area.

When asked in item AQ13f whether guidance material on the topic of language testing and assessment was very user-friendly, a question intended to investigate concerns similar to those raised by Taylor, (2009, p.21) who suggests that some teachers engaged in the process of language testing may feel that it is an overly complex practice which is accessible only to experts, only 25.3% of respondents agreed or strongly agreed.

Item AQ13g was designed in order to ascertain the extent to which respondents felt that they had received sufficient training in EAP testing and assessment. This is a question directly linked to

concerns expressed by Sharpling (2002) and suggestions connected to the quality of training provision related to EAP and assessment (Brown and Bailey, 2008). Notably only 36.7% (58) respondents agreed or strongly agreed that their training had been sufficient.

AQ13h was created in order to gauge respondent understanding of the relative complexity of EAP assessment in comparison to the assessment of subject knowledge, a question which addressed the challenges and importance of ensuring of construct validity in EAP assessment (Blue et al., 2000, p.26-27). The results of this question revealed that (92) 58.2 % answered either 'Disagree' or 'Strongly Disagree' and thus confirmed their understanding of the relative complexity of EAP assessment.

When asked in item AQ13i whether it is straightforward to identify the key EAP language areas and skills that students need to be assessed in, only 36% of respondents agreed or strongly agreed. Finally in this grouping, when in AQ13j respondents were asked if they find it straightforward to identify a question which is not working properly, a question probing means of interpreting tests and methods, which could include statistical procedures, used for analysing complex data and identifying trends (Woods et al., 1986, p.1). Notably, only 38.6 % (61) of respondents agreed or strongly agreed.

4.3.2.1.2 Responses to items AQ14-A-M

Item AQ14 contained 13 sub items A-M which were created in order to allow respondents to rate their level of skill in various aspects of EAP Assessment Literacy. These areas were drawn from the Test Design Framework (Fulcher, 2010, p.94), key aspects of Assessment Literacy, as identified by Stiggins (1995) and Popham, (2001), along with the context of contemporary EAP as

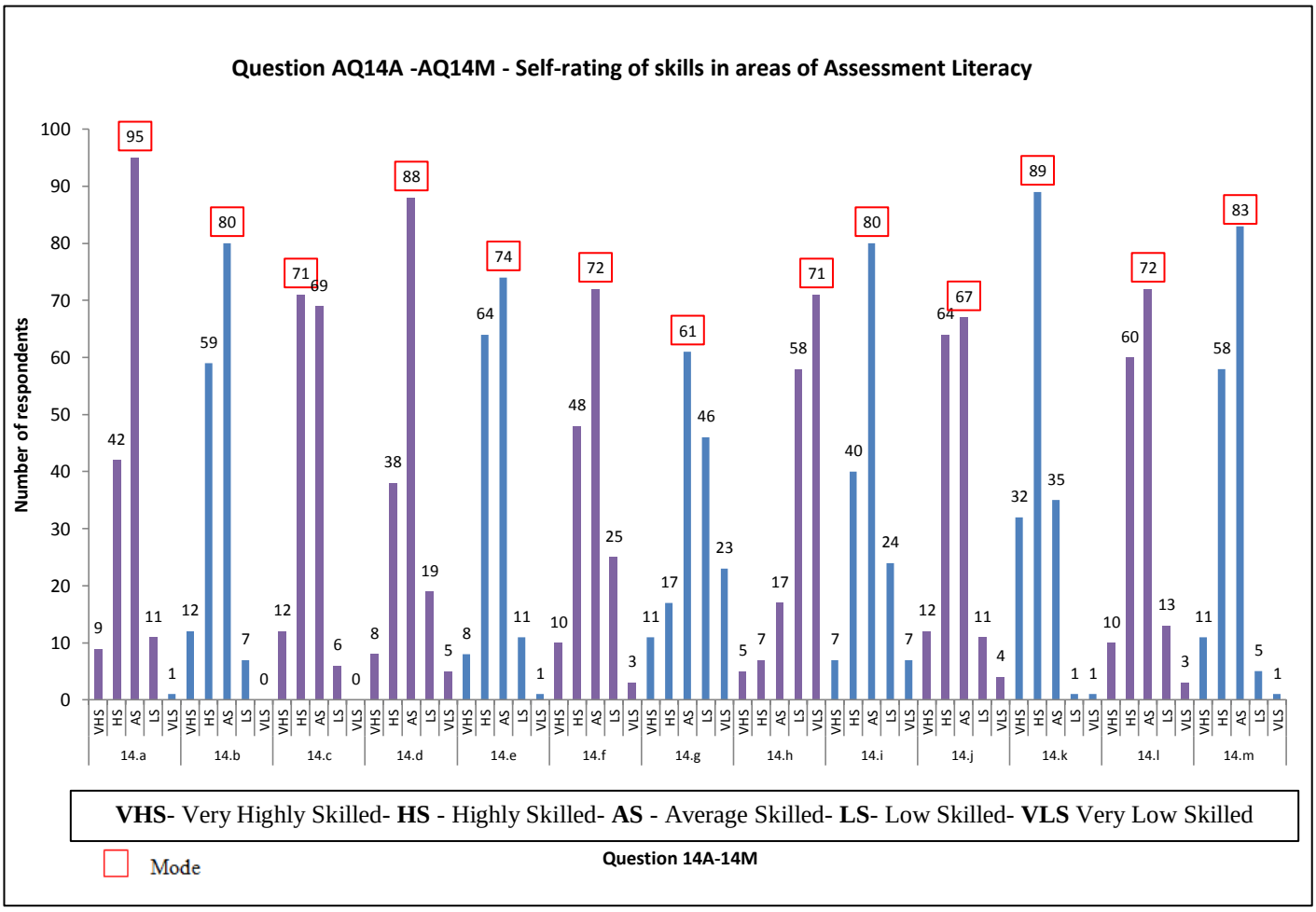
described by experts in the field such as Alexander et al. (2008). Table 53 below displays the questions which were posed:

Table 53 Items/Questions in set AQ14

Item code	Item
AQ14a	Please rate your skills in test and assessment design
AQ14b	Please rate your skills in defining test purpose
AQ14c	Please rate your skills in identifying what should be assessed
AQ14d	Please rate your skills in trialling/piloting tests
AQ14e	Please rate your skills in writing test items/questions
AQ14f	Please rate your skills in developing and using grading procedures
AQ14g	Please rate your skills in using descriptive statistics (such as mean, mode and median) to analyse test and assessment scores
AQ14h	Please rate your skills in using inferential statistics (such as Cronbach's Alpha, T-tests and factor analyses)
AQ14i	Please rate your skills in drawing inferences from test and assessment scores
AQ14j	Please rate your skills in making instructional decisions based on test scores
AQ14k	Please rate your skills in giving feedback to students based on test scores
AQ14l	Please rate your skills in helping other people who use EAP test results to understand them
AQ14m	Please rate your skills in identifying problems with tests or assessments

Chart 7 below displays the frequency of responses from items AQ14A-M.

Chart 7 Questions AQ14A-AQ14M for population B



With regard to test and assessment design, a key aspect of ensuring test usefulness (Bachman and Palmer, 1996, p.38) AQ14a revealed that only 32.2 % (51) rated themselves as highly skilled or very highly skilled in this area. AQ14b invited respondents to rate their skills in defining test purpose, a critical quality in EAP assessment, as supported by professional bodies and experts in the EAP field (BALEAP, 2012a; Schmitt, 2012) to which only 8.2% (13) of respondents rated themselves as highly skilled or very highly skilled. As far as identifying what should be assessed is concerned, a challenge acknowledged by Kane in connection with domain sampling (2012, p.41-42), 52.5% (83) of respondents to AQ14c indicated that they were highly skilled or very highly skilled in this area. In the area of trialling/piloting tests, which is closely connected to the area of Assessment Literacy associated with the prevention of test-related problems (Davies et al., 1999, p.150), item AQ14d shows that 29.1% (46) of respondents believed themselves to be highly skilled or very highly skilled in this area. In connection with test item writing AQ14e revealed that 45.5% (72) of respondents consider themselves to be highly skilled or very highly skilled in this respect. On the matter of developing and using grading procedures, an important matter linked to an outcomes-based approach (Fry et al., 2003, p.26-41) providing evidence of a link between scores, construct and domain (Fulcher and Davidson, 2007 p.91-114), 36.7% (58) of respondents to item AQ14f claimed to be either highly skilled or very highly skilled.

Questions AQ14g and AQ14h sought to measure respondents' perceptions regarding their skill levels in using descriptive statistics and inferential statistics, which allow findings to be explored more widely than simply within the confines of individual samples (Dörnyei, 2003, p.115; Larson-Hall, 2010, p.45). Results show that only 17% (28) of respondents consider themselves to be highly skilled or very highly skilled in the use of descriptive statistics and 7.5% (12) respondents believed themselves to be highly skilled or very highly skilled in using inferential statistical procedures, which can explore beyond the scope of descriptive methods (Brown,1988, p.115).

With regard to respondents' perceptions connected to their ability to draw inferences from test and assessment scores, a crucial aspect of construct validity (Messick, 1989), 29.7% (47) of respondents to AQ14i identified themselves as highly skilled or very highly skilled. 48% (76) of respondents reported in AQ14j that they were either highly skilled or very highly skilled in making instructional decisions based on test scores. Items AQ14k and AQ14l relate to respondent skill in giving feedback to students based on test scores and helping other individuals who use test scores to understand them. This is both a crucial aspect of Assessment for Learning (Biggs & Tang, 2011, p.64; Knight, 1995), acknowledged as crucial to Assessment Literacy (Popham, 2009, p.11), and a key aspect of the education of test score users. The results show that (121) 76% of respondents feel either skilled or highly skilled in giving feedback to students and (70) 46% of respondents feel highly skilled or very highly skilled in helping other stakeholders to interpret test scores.

Finally, when considering respondents' skill in identifying problems with EAP tests or assessments, a key aspect of the proactive call to action, as identified by Popham (2001), 43% (69) rate themselves as highly skilled or very highly skilled.

When considering the modal response to each of the items in Chart 7, it can be seen that 9 of the 13 items show a clear modal response in the 'Average Skilled' category with the modes ranging from 61 respondents (38.6%), in response to AQ14g, to 95 respondents (60.1%) in item AQ14a. Items AQ14c and AQ14j show a bimodal response across 'Average Skilled' and 'Highly Skilled' categories. AQ14c shows 69 (43.6%) respondents feel 'Average Skilled' whilst 71 (44.9%) respondents feel highly skilled. AQ14j shows that 67 (42.4%) respondents claim to be 'Average Skilled' whilst 64 (40.5%) respondents report to be highly skilled. The mode for item AQ14k

shows a mode of 89 (56.3%) respondents in the 'Highly Skilled' band. The only mode appearing in the very 'Low Skilled' band was in item AQ14h which showed at total of 71 (44.9%) of respondents. Table 54 can also be used to show the ranking of respondents according self-evaluations of 'High Skill' or 'Very High Skill' level in each of the skill areas represented in this set of items.

Table 54 Ranking of respondent skill areas in EAP assessment according to perceptions of high or very high skill

Item code	Respondent skill area in EAP assessment	Percentage of respondents claiming to be highly skilled or very highly skilled
AQ14k	skills in giving feedback to students based on test scores	76
AQ14c	skills in identifying what should be assessed	52.5
AQ14j	skills in making instructional decisions based on test scores	48
AQ14l	skills in helping other people who use EAP test results to understand them	46
AQ14e	skills in writing test items/questions	45.5
AQ14m	skills in identifying problems with tests or assessments	43
AQ14f	skills in developing and using grading procedures	36.7
AQ14a	skills in test and assessment design	32.2
AQ14i	skills in drawing inferences from test and assessment scores	29.7
AQ14d	skills in trialling/piloting tests	29.1
AQ14g	skills in using descriptive statistics	17
AQ14b	skills in defining test purpose	8.2
AQ14h	skills in using inferential statistics	7.5

4.3.2.1 Responses to items linked to approaches to EAP testing and assessment design

4.3.2.1.1 Responses to items AQ16-A-J

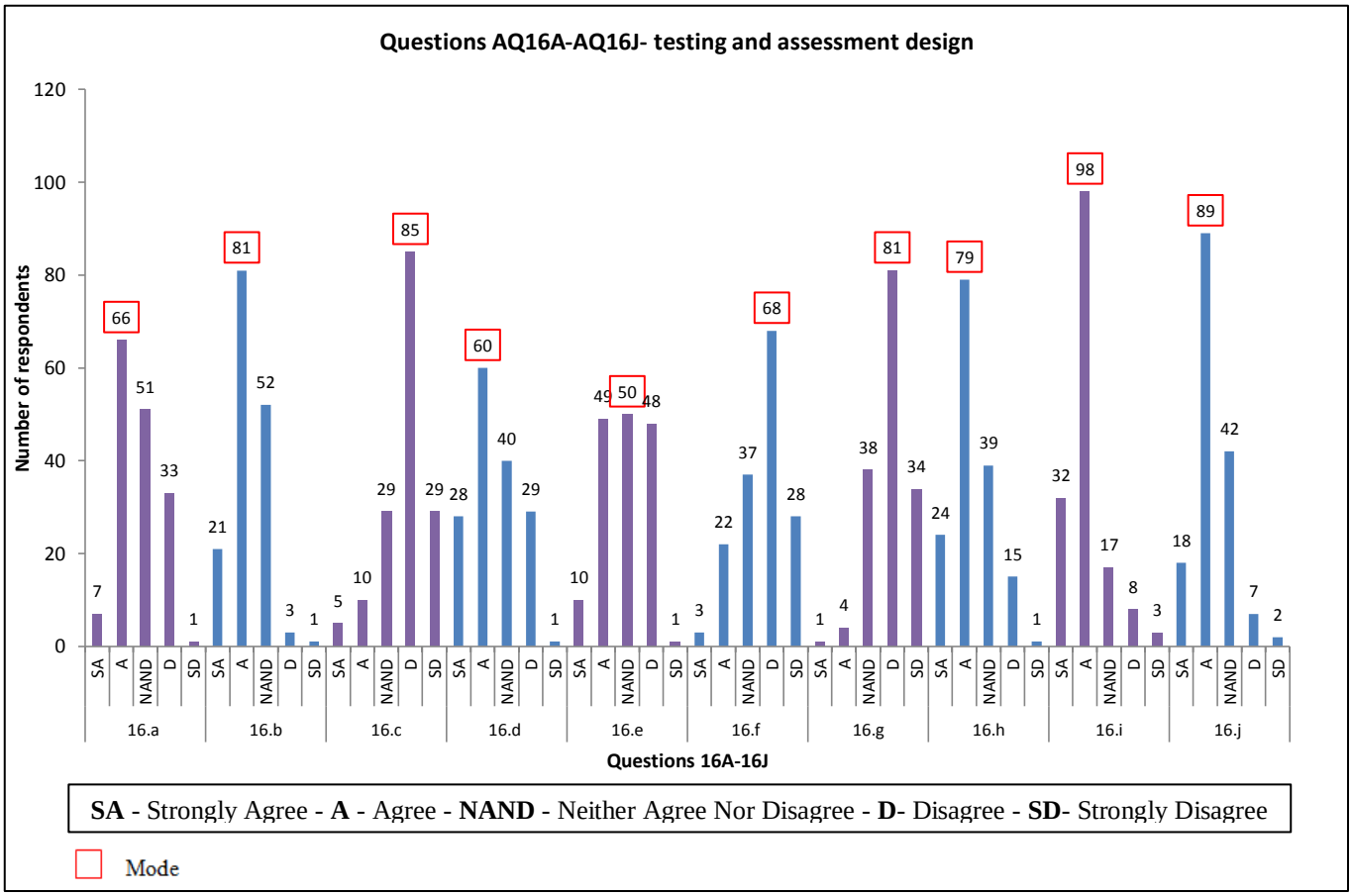
Item AQ16 contained 10 sub items A-J which were produced in order to question respondents on their views concerning EAP test and assessment design. Table 55 below displays the questions which were posed:

Table 55 Items/Questions in set AQ16

Item code	Item
AQ16a	With practice, EAP test design can become a straightforward process.
AQ16b	It is necessary to use a test specification when designing a test.
AQ16c	It is unhelpful to involve academic content tutors in the process of EAP test design.
AQ16d	When designing an EAP test or assessment, it is necessary to take into account the specific subjects that students will be studying.
AQ16e	Each EAP testing situation will require its own bespoke test.
AQ16f	A good starting point in EAP testing is to write the test questions.
AQ16g	EAP test design is best left to the individual class teacher.
AQ16h	Each different testing and assessment situation may need its own marking descriptors.
AQ16i	It would be appropriate to design an EAP diagnostic test, in order to measure which students can use English effectively for university study.
AQ16j	An EAP achievement test would be a suitable test to design in order to evaluate a learner's understanding of a specific EAP skill.

Chart 8 below displays the frequency of responses from items AQ16A-J.

Chart 8 Questions AQ16A-AQ16J for population B



The results of question AQ16a showed that (73) 46.2% of respondents indicated agreement or strong agreement with the suggestion that the design process can become more straightforward

with practice, an indication of the inherent complexity of developing assessments in EAP (Douglas, 2000, p.ix) and the iterative nature of test design (Fulcher and Davidson 2007, p.61).

In addition, in response to AQ16b (102) 64.5 % of respondents reported agreement or strong agreement with the need to use a test specification as a principled guidance tool (Davidson and Lynch, 2002). With respect to the involvement of academic content tutors in the process of EAP test design and identifying students' communicative needs (Weir, 1983, p.547), AQ16c demonstrated that 72.1% (114) of respondents disagreed or strongly disagreed that the involvement of academic content tutors was unhelpful.

Respondents' views on the need to consider the specific subjects which students are studying in the process of EAP test and assessment design (see reference to ESAP/EGAP debates as referenced in Hyland, 2006), was dealt with in AQ16d, in response to which 55.6% (88) of respondents either disagreed or strongly disagreed. The collection of respondents' perceptions concerning the need for bespoke tests to fit with different testing situations was the aim of AQ16e (see challenges suggested by Kane, 2012, p.41-42). Notably, the results of the item revealed a multi-modal response with 'Agree', 'Neither Agree Nor Disagree' and 'Disagree' returning 31.0% (49), 31.6% (50) and 30.3% (47) respectively.

Item AQ16f was created in order to ask respondents about their views on a suitable starting point for EAP testing and assessment, in light of research good practice such as that recommended by the Test Design Framework (Fulcher, 2010, p.94). In response to this item, 60.7% (96) of respondents strongly disagreed that writing the test items represents a good point to start in the EAP test design process. In question AQ16g, as a measure of respondents' understanding of the importance of teamwork and consensus in test development (Fulcher & Davidson 2007, p.61),

respondents were asked to comment on whether EAP test design is best left individual class teachers. The responses showed that 72.7 % (115) of respondents disagreed or strongly disagreed with this concept.

Item AQ16h was used to investigate the extent to which respondents feel that different testing and assessment situations may need their own marking descriptors, a reference both to outcomes based assessment (Bloom, 1956) and the concept of indigenous assessment criteria, related to specific domains (Douglas, 2000). Responses to this question showed that 65.1% (103) of respondents agreed or strongly agreed with the need to use different marking descriptors for different testing situations. Both questions AQ16i and AQ16j were intended to check respondents' understanding of the situations when it is appropriate to use different test types.

Responses to AQ16i showed that 82.2 % of (130) respondents agreed or strongly agreed that an EAP diagnostic test would be suitable for use to measure which students can use English effectively for university study. When asked in AQ16j whether an achievement test would be a suitable test to design in order to evaluate a learner's understanding of a specific EAP skill, 67% (107) of respondents either agreed or strongly disagreed.

4.3.2.1.2 Rankings relating to key stages of test design (AQ17)

Items 1-8 in AQ17 asked respondents to rank stages of EAP assessment design and development, inspired by aspects of the Test Design Framework (Fulcher, 2010, p.94) and key factors in Assessment Literacy (Stiggins, 1995 and Popham, 2001). The aim was to attempt to determine aspects of teachers' Assessment Literacy through observing their understanding of key processes involved in test design and development and by choosing from a list of nine options, thus forcing the exclusion of one item. Whilst it is not inferred that one fixed process exists, the researcher wished to identify what respondents prioritise.

The options for selection which correspond to options A-I are as follows:

A. Create the questions/items:

B. Piloting and field testing:

C. Research the skills which require assessment:

D. Identify the sub skills/abilities to be tested (Constructs):

E. Create or consult a test specification:

F. Define the test purpose:

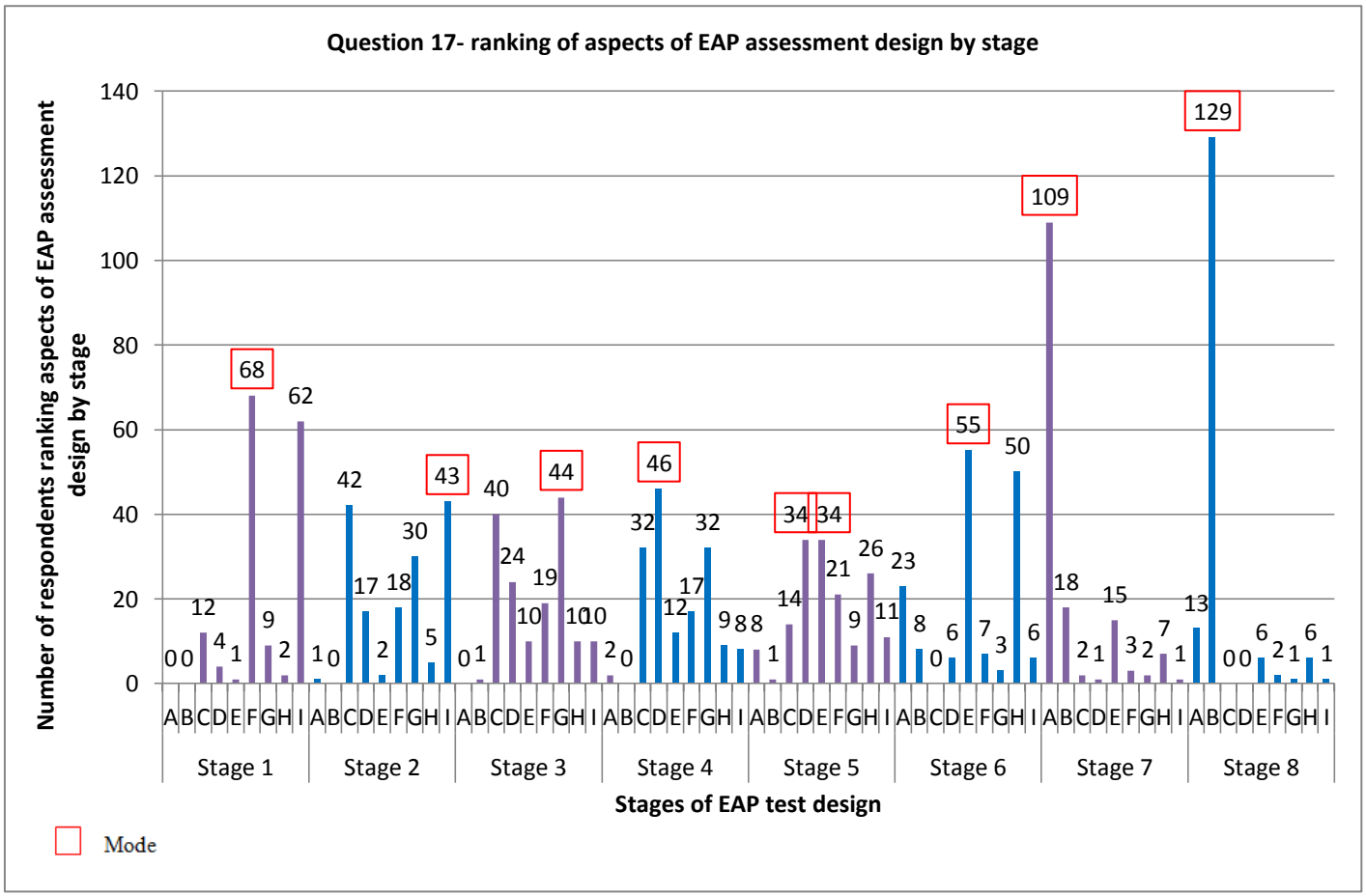
G. Define main skills to be tested (reading, writing, listening, speaking etc.)

H. Consult past papers or versions of the test:

I. Investigate the subject area(s) where students will use EAP the subject area(s) where students will use EAP

Consequently, when presenting the results of this section, findings can be described in terms of frequency of response to each individual item as well as frequency when each of the eight stages are summed, in order to reveal which items were most frequently selected or excluded. Chart 9 below shows the responses to item AQ17:

Chart 10 Question AQ17 for population B



The modal response for Stage 1 of the test design process was identified by 43% (68) of respondents as option F, who suggested that 'Defining test purpose', an important aspect of EAP test design (Alexander et al., 2008), should be the first stage of test and assessment design. This figure was closely followed by 39.2% (62) of respondents who suggested that investigating the subject area where students will use EAP could be an alternative first stage. This left only 17% of respondents selecting an alternative option.

For Stage 2 of the process, both options C, 'Research the skills which require assessment' and option I, 'Investigate the subject area where students will use EAP', received similarly high rates of response with I gaining 27.2% (43) and C gaining 26.5 (43). The next highest response rate was

for option G, relating to the definition of the main skills to be tested, such as reading, and writing, listening and speaking 18.9% (30).

At stage three of the process the modal response identifies item G as the most popular with 29.1% (46) of respondents, which suggests that the definition of the main skills to be tested was the most popular choice for this stage. This is closely followed by item C, 'Research the skills which require assessment.'

There is a clearer modal response for Stage 4 which shows that 29.1% (46) of respondents believe that this stage should involve option D and 'Identification of the sub skills/abilities to be tested.' Options C and G are then the next most frequently selected items, each with 18.9% (30) of respondents.

By Stage 5 of the process there appears to be a less distinct set of responses with items D and E showing a bimodal response with D related to the definition of skills that are to be tested and E, referring to the need to 'Create or consult a test specification'.

Stage 6 shows a greater consensus of opinion with the modal response being option E at 32.8% (55), and the need to 'Create or consult a test specification', a key guidance tool in test development (Davidson & Lynch, 2002). The next response with the highest number of respondents was item H with 31.6% (50) of respondents, which involves the consultation of past papers or versions of the test, in recognition of reverse engineering where test development often commences (Fulcher &Davidson, 2007, p.57). At 14.5% (23) of the total number of respondents, Item A, 'Create the questions/items' was the third most popular response.

Stage 7 returned a strong modal response with 68.9% (109) of responses identifying that this stage required the creation of question or test items. Items B and E were the next most popular responses with 11.3% (18) and 9.4% (15) of respondents. Items B and E relate to piloting and the need to create or consult a test specification respectively.

Stage 8 received the highest modal response of all items with 81.6% (129) of respondents identifying 'Piloting and field testing', a part of pre-testing (Bachman & Palmer, 1996, p.234) as the purpose of this stage. The next most frequently selected option was item A, relating to the creation of test questions.

The composite results of each of the 8 stages included in item AQ17 can assist in identifying which of the 9 options provided were the most frequently included and excluded within the 8 stages available to respondents. Whilst the researcher acknowledges that in real EAP practice, assessors are not restricted to eight stages, this process forces respondents to show which options they prioritise.

Table 56 Composite results of stages 1-8 drawn from AQ17

Option	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5	Stage 6	Stage 7	Stage 8	Total
A	0	1	0	2	8	23	109	13	156
B	0	0	1	0	1	8	18	129	157
C	12	42	40	32	14	0	2	0	142
D	4	17	24	46	34	6	1	0	132
E	1	2	10	12	34	55	15	6	135
F	68	18	19	17	21	7	3	2	155
G	9	30	44	32	9	3	2	1	130
H	2	5	10	9	26	50	7	6	115
I	62	43	10	8	11	6	1	1	142

Table 56 above shows the eight stages of test design and the nine options made available to respondents. The results show that the options most frequently included at some point within the eight stages were options B, A and F, relating to piloting, item creation and the definition of test purpose. In contrast, item H, relating to the consultation of past papers or versions of a test, was the most frequently omitted option from the stages of test design with this having been omitted by 43 respondents. Items G, D and E, were the series of items with the next highest incidence of omission. These items relate to the identification of main skills, sub skills and the creation or consultation of a test specification. Item G was omitted by 28 respondents, Item D was left out by 26 respondents and item E by 23 respondents.

4.3.2.2 Responses to items linked to views on/ experience of validity and reliability in EAP testing and assessment

4.3.2.2.1 Responses to items AQ19-A-J

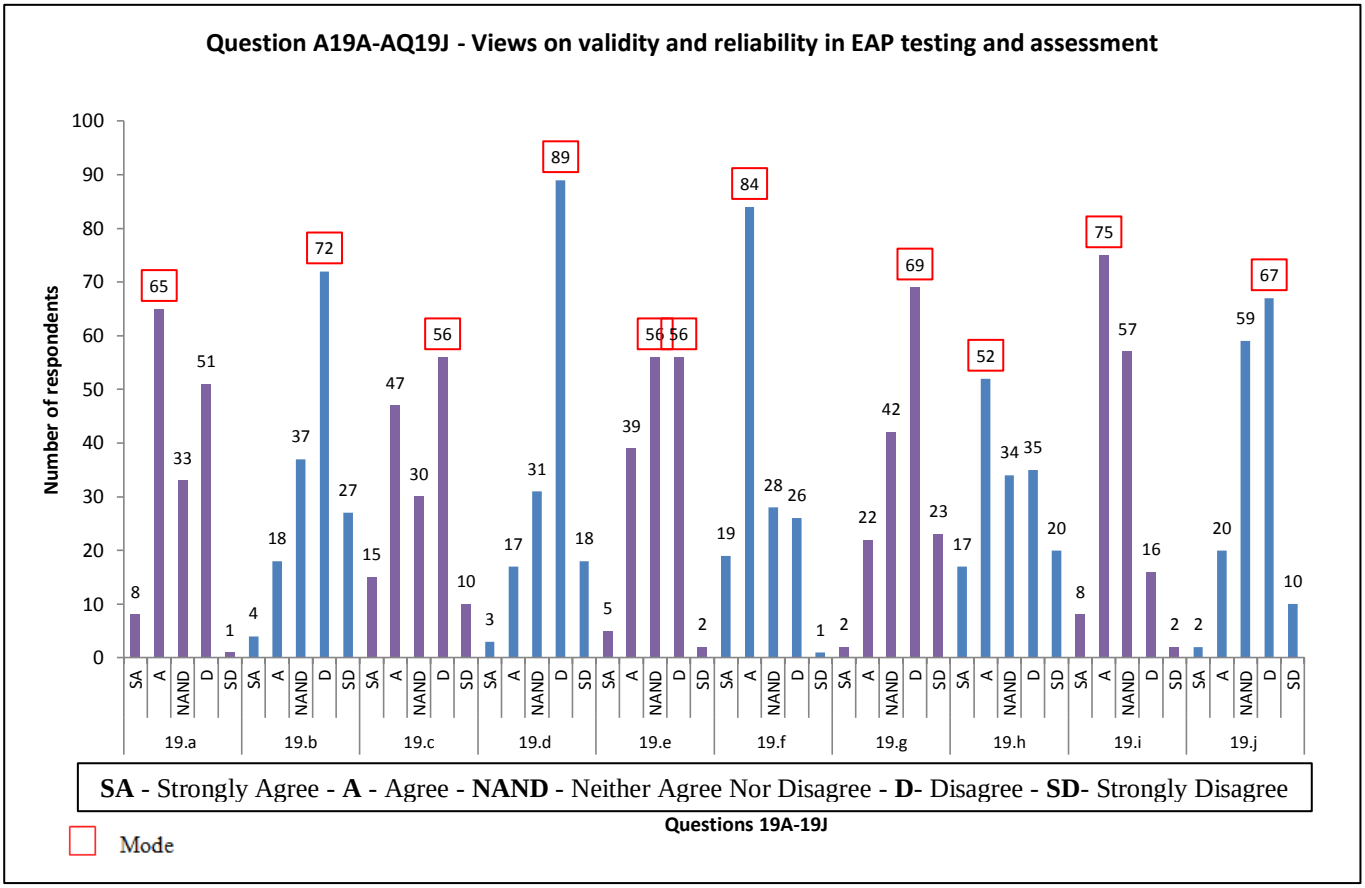
Item AQ19 contained 10 sub items A-J which were produced in order to question respondents on their views concerning validity and reliability in EAP testing and assessment, which represent key aspects of Assessment Literacy. Table 57 below displays the questions which were posed:

Table 57 Items/Questions in set AQ19

Item code	Item
AQ19a	It is difficult to be specific about what constitutes a particular area of linguistic skill in EAP.
AQ19b	Decisions based on test results from IELTS or TOEFL are more likely to be valid and reliable than those based on the results of my own EAP assessments.
AQ19c	EAP tests should use reading and listening texts created specifically for EAP language testing
AQ19d	When creating EAP tests, it is unreasonable to expect EAP tutors to use language from the different academic fields in which students' English will be applied.
AQ19e	When selecting EAP writing exercises and assessments for students preparing to study Mathematics, it would be best to avoid essays.
AQ19f	Students studying different academic subjects are likely to need different types of EAP assessment.
AQ19g	When a test is well compiled, a single question can be an effective way to measure any given language ability.
AQ19h	A reliable EAP test will also be valid.
AQ19i	Giving students a different version of one of my EAP tests, aimed at assessing the same skills, would likely result in students getting a similar mark.
AQ19j	A class of students gaining a low grade in an EAP assessment is a good indication that the assessment mechanism was faulty in some way.

Chart 11 below reports the frequency of responses from items AQ19A-J.

Chart 11 Questions AQ19A-AQ19J for population B



The results of AQ19a, intended to investigate respondents' views on the complex and time-consuming stage of construct definition (Brown & Abeywickrama, 2010, p.108), show that 46.2% (73) agree or strongly agree that it is difficult to be specific about what constitutes a particular area of linguistic skill in EAP, this leaves 32.9% (52) of respondents who disagree or strongly disagree. Views associated with standardised EAP-related tests were probed in item AQ19b with regard to confidence in the validity of test results from systems such as IELTS or TOEFL (Alexander et al., 2008, p.308). Results show that 62.6 % (99) of respondents disagree or strongly disagree that results from IELTS or TOEFL are more likely to be valid and reliable than those based on the results of my own EAP assessments. AQ19c attempts to question respondents regarding the importance they attach to using authentic/real academic tests in EAP language testing. Whilst the modal response lies in the 'Disagree' field with 35.4 % of (56), overall the main pattern shows that respondents are split relatively evenly between 'Strongly Agree' and 'Agree' at 39.2% (62) and disagree and strongly disagree at 41.7% (66).

When in AQ19d respondents were asked if it was unreasonable to expect EAP tutors to use language from the different academic fields in which students' English will be applied, a view espoused by some EAP researchers such as Spack, (1988), the responses showed that 67.7 % (107) of individuals either disagreed or strongly disagreed. The number of respondents who selected the 'disagree' response for this question constituted the highest modal response amongst each of the 10 sub items in this series of questions.

Question AQ19e aimed to determine respondent views associated with construct validity (Mesick, 1989) and the use of question types in tests which reflect tasks from the students' target academic domain. In response to this question, which invited respondents to consider, when EAP writing exercises and assessments for students preparing to study Mathematics, if it would be best to avoid

essays, the frequency of responses demonstrated that only 27.8% (44) either agreed or strongly agreed. In fact, the data demonstrated a bimodal frequency with 35.4 % (56) of respondents selecting each of the 'Neither Agree Nor Disagree' and 'Disagree' options.

65.1% (103) of respondents either agreed or strongly agreed in response to AQ19f when they were asked to comment on whether students studying different academic subjects are likely to need different types of EAP assessment which reflect target language use (Douglas, 2000, p.103).

Item AQ19g was designed to probe teachers' understanding of the unreliability of single measures of a construct in a test compared to multiple measures (Fulcher, 2010, p.57). The distribution of responses to this item indicates that 58.2% (92) of individuals disagree or strongly disagree with the effectiveness of a single question in measuring language ability.

In order to examine whether respondents understand that reliability does not infer validity (Bachman, 1990, p.241), question AQ19h was posed. Although 34.8% (55) of respondents disagreed or strongly disagreed with the fact that a reliable EAP test will also be valid, this left 43.67% (69) of respondents who strongly agreed or agreed.

With regard to respondents' confidence in the reliability of their own EAP language tests and assessments, a concern expressed by Coniam (2009), the results on AQ19i showed that 83.5% (132) of respondents were either in agreement or strong agreement that their use of different test versions returns results which are reliable.

Finally, in AQ19j, in order to investigate whether respondents understand that poor test scores can highlight poor tests as well as poor test-taker proficiency (Shohamy, 2001), when asked if a class

of students gaining a low grade in an EAP assessment could be considered a good indication that the assessment mechanism was faulty in some way, responses showed 48.7% (77) of respondents either disagreed or strongly disagreed whilst only 13.9% of respondents (22) either agreed or strongly agreed. This left 37.3% (59) of respondents neither agreeing nor disagreeing.

4.3.2.2.2 Responses to items AQ20-A-H

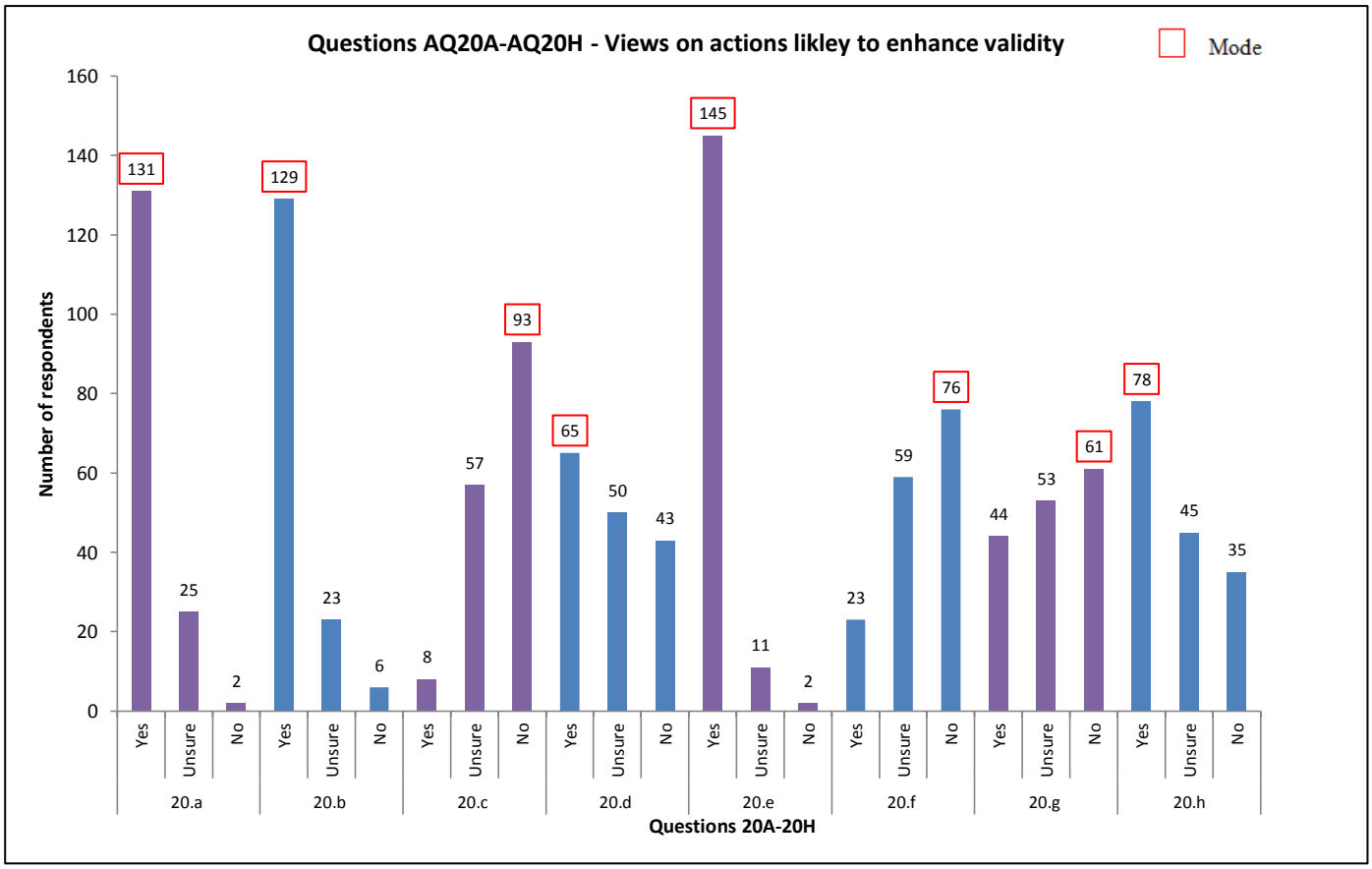
Item AQ20 contained 8 sub items A-H which were produced in order to collect views from respondents regarding actions which are likely to enhance validity of inferences drawn from test scores, a key aspect of construct validity in practice, as supported by recent texts created by practice-oriented researchers such as Fulcher (2010). Table 58 below displays the questions which were posed:

Table 58 Items/Questions in set AQ20

Item code	Item
AQ20a	Researching the subject contexts in which students are required to use EAP
AQ20b	Incorporating multiple measures of EAP language abilities
AQ20c	Adding more multiple choice EAP questions
AQ20d	Requiring all EAP tutors in a team to mark all test sections
AQ20e	Working as a team with other EAP colleagues to determine what should be tested
AQ20f	Allowing EAP markers to use their own judgement when grading
AQ20g	Phrasing EAP questions so that they sound academic
AQ20h	Including a question for each area of EAP which has been taught in the course

Chart 12 below reports the frequency of responses from items AQ20a-h.

Chart 12 Questions AQ20A-AQ20H for population B



Question AQ20a was introduced in order to examine whether respondents feel that inferences drawn from tests are likely to be considered more valid if tests are based on research into subjects which students are using EAP in. These challenges are discussed by Kane (2012). The results showed that 82.9% (131) of respondents answered 'Yes' to this item.

With reliability and the multifaceted nature of many language constructs in mind, Question AQ20b was introduced, in order to determine if respondents felt that incorporating multiple measures of a construct would enhance the validity of inferences drawn from test scores (Fulcher and Davidson, 2007, p.105). Once again the results were pronounced with 81.6% (129) of respondents selecting 'Yes' and confirming that adding multiple measures is likely to improve the trustworthiness of inferences drawn from tests.

In response to AQ20c 58.8 % (93) of respondents disagreed that just by adding more multiple choice questions the validity of inferences from test scores could be improved. This however left 36% (57) of respondents were unsure and 5% (8) who agreed.

Item AQ20d was designed to explore views on inter-rater reliability, a key tool in maintaining reliability in test scoring (Brown & Abeywickrama, 2010, p.41-42), and the problems associated with validity of inferences drawn from tests which can arise if there is a lack of consistency in the approach to test marking. When asked if all EAP tutors in a team should be asked to mark all test sections, the results revealed that the modal response was 'Yes' with 41.1% (65) of respondents selecting this option. 31.6% (50) of respondents selected the 'Unsure' option. This left 27.2% (43) respondents reporting 'No'.

The importance of consensus (Fulcher and Davidson 2007, p.61) through team work was highlighted in item AQ20e. Respondents were asked if they felt that working as a team with other EAP colleagues, to determine what should be tested, could enhance the validity of inferences drawn from tests. The findings for this item reveal that 91.8% (145) of respondents agree that this activity is likely to enhance the validity of inferences drawn from tests or assessments.

Question AQ20f was included to determine the extent to which respondents believe the validity of inferences drawn from test scores can be enhanced through the use of marker personal judgement during the marking process, a question designed to probe the use of criterion referencing and grading criteria (McNamara, 2000, p.64). Whilst the modal response was 'No' with 48.1% (76) of respondents choosing this option, the remaining respondents opted to either record a neutral central position at 37.3 % (59) or agreement at 14.6% (23).

In response to item AQ20g, 38.6% (61) of respondents disagreed that just phrasing a question so that it sounds more academic, a process which could be considered a limited and impressionistic aspect of face validity (Henning, 1987, p.96), would enhance the trustworthy nature of inferences drawn from test scores. This left 33.5% (53) of respondents who were unsure and 27.8% (44) who agreed.

Finally, item AQ20h sought to question respondents on the topic of content validity by asking if they felt that the validity of inferences drawn from test scores could be enhanced if the test was designed to include questions for each area of EAP which has been taught in the course. The results showed that 39.4% (78) of respondents answered 'Yes', leaving 28.4% (45) unsure and 22.1% (35) answering 'No'.

4.3.2.3 Responses to items linked to practices involving analysis and interpretation of the results and scores of EAP tests and assessments

4.3.2.3.1 Responses to items AQ22-A-J

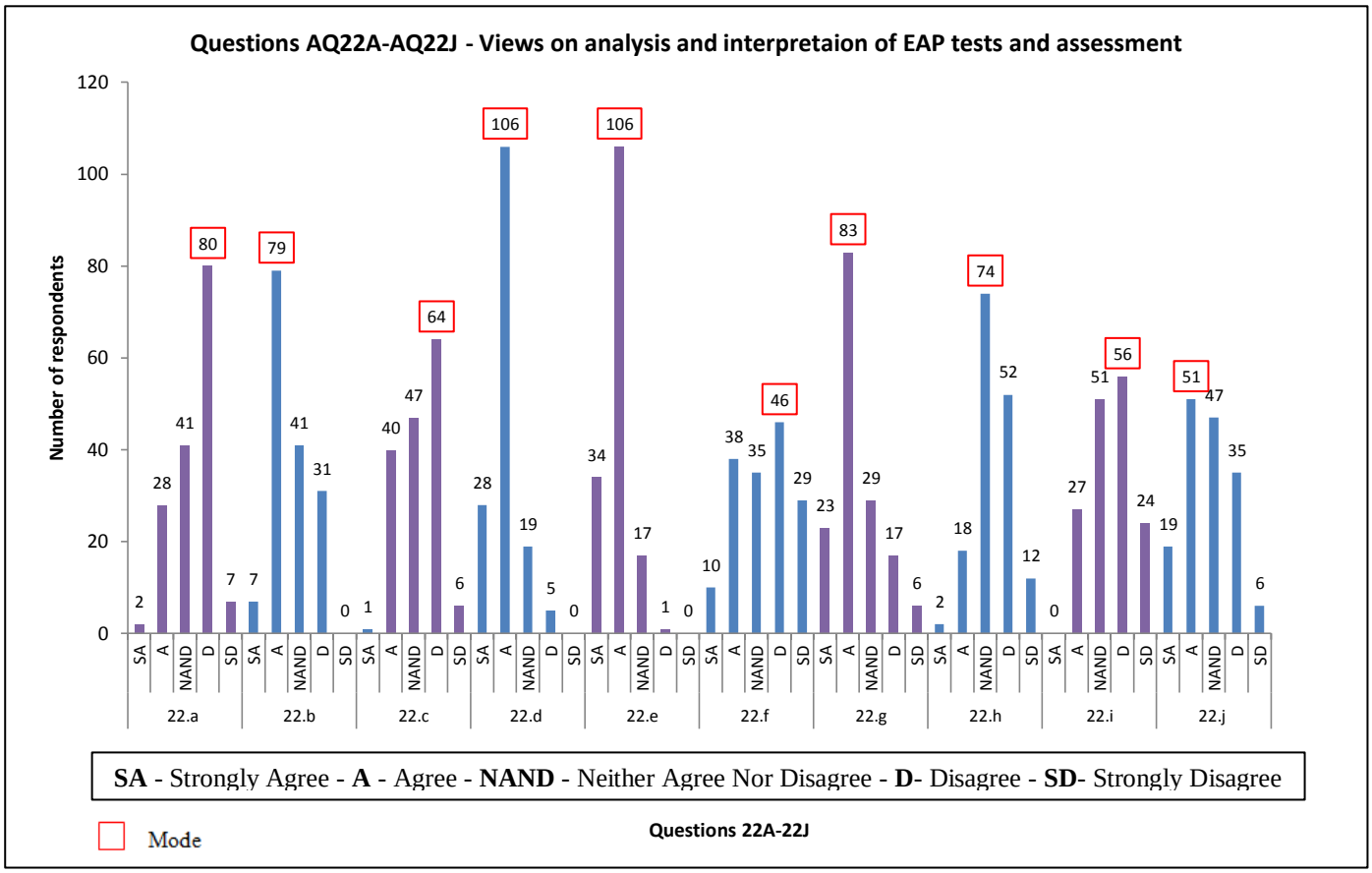
Item AQ22 contained 10 sub items A-J which were produced in order to collect views from respondents on situations concerning test and assessment analysis and interpretation. Table 59 below displays the questions which were posed:

Table 59 Items/Questions in set AQ22

Item code	Item
AQ22a	Analysing the results of EAP tests and assessments is a relatively straightforward process
AQ22b	It is really difficult to draw concrete conclusions from EAP test and assessment results alone.
AQ22c	The results from one good EAP test can generally provide a very accurate overall impression of a student's breadth of ability.
AQ22d	EAP students who are required to reach a certain level of ability in their EAP skills should be evaluated in relation to specific learning outcomes (criterion referencing).
AQ22e	The results of EAP tests should be explained in more depth so that other colleagues can understand what skills they relate to.
AQ22f	I am frequently required to help academics in other departments to understand students' EAP grades more effectively.
AQ22g	The EAP department is responsible for making sure that users of the test results interpret their meaning appropriately.
AQ22h	A class of 17 students all gain low scores in section 3 of an EAP test. This is most likely to mean that the students need to develop their skills in that area further.
AQ22i	EAP students who need to demonstrate a particular skill level ought to be assessed and selected based on their rank compared to other students who completed the test (norm referencing).
AQ22j	It is impractical to convert EAP test results into IELTS or TOEFL grades as the process is too complex.

Chart 13 below presents the range of responses provided through items AQ22A-J.

Chart 13 Questions AQ22A-AQ22J for population B



The first question in this series, AQ22a, was designed to collect information regarding respondents' views on the complexity of interpreting test scores, the challenges of which are described by Brown and Abeywickrama, (2010, p.285). Notably, 'Disagree' represented the modal response with 50.6% (80) of respondents. In total, 55% (87) of respondents either disagreed or strongly disagreed with the statement that analysing EAP tests and assessments is straightforward and only 19% (30) of respondents agreed or strongly agreed. This left 25.9% (41) of respondents in the 'Unsure' category.

Item AQ22b asked respondents to determine the extent to which they find it difficult to draw concrete conclusions from EAP test and assessment results alone. In an almost identical pattern to AQ22a, the results of AQ22b show that 54.4% of (86) respondents either agreed or strongly agreed whilst 19.6% (31) disagreed and 25.9% (41) of respondents were unsure.

With regard to the reliability of test score interpretations from a single EAP test, the responses to AQ22c demonstrated that 44.3% (70) of respondents disagree or strongly disagree that one good EAP test can generally provide a very accurate overall impression of a student's breadth of ability. The modal response was 40.5% (64) in the 'Disagree' category. Notably only (41) 25.9% of respondents agreed or strongly agreed that a single EAP test could provide an accurate impression of a student's breadth of ability. This left 29.7% (47) of respondents who neither agreed nor disagreed.

AQ22d was designed to measure respondents' understanding of testing situations when it is appropriate to use criterion referencing (Henning, 1987, p.4-8). The results showed that 84.8% (134) of respondents either agreed or strongly agreed that EAP students who are required to reach a certain level of ability in their EAP skills should be evaluated in relation to specific learning

outcomes. In this case, the modal response showed that 67% (106) selected the ‘Agree’ category. This left only 12% (19) of respondents who were unable to agree or disagree and 3.1% (5) who disagreed.

AQ22e also related to outcomes and criterion referencing (Bloom, 1956; Fry et al., 2003, p.26-41) in terms of creating meaningful explanations of EAP test scores so that other stakeholders can understand what they mean. The results of this item demonstrate that 88.6% (140) of respondents agree or strongly agree that EAP test results should be explained in more depth so that other colleagues can understand what skills they relate to. The modal response was 67% (106), the same as in AQ22d. This left only 10.8% (17) of respondents who were unable to agree or disagree and 0.6% (1) who disagreed.

When in AQ22f respondents were asked to comment on whether they are frequently required to help academics in other departments to understand students' EAP grades more effectively, an important part of stakeholder education in Assessment Literacy (Stiggins, 1995), the frequency of responses was less well defined. Although 30.3 % (48) of respondents either agreed or strongly agreed, the modal response was ‘disagree’ at 29.1% (46) and in total 47.4% (75) of respondents either disagreed or strongly disagreed.

With regard to respondents’ beliefs about the EAP department having responsibility for making sure that users of the test results interpret their meaning appropriately, the results of AQ22g indicate that 67% (106) of respondents either agreed or strongly agreed. Only 14.6% (23) of respondents disagreed or strongly disagreed with this view, leaving 18.3% (29) who neither agreed nor disagreed.

In AQ22h, in response to a situation, surrounding a class of students who all gained a low score in a particular section of an EAP test, 12.7% (20) of respondents either agreed or strongly agreed that this this is most likely to mean that the students in question need to develop their skills in that area of further. 40.5% (64) either disagreed or strongly disagreed with the suggested stance, leaving the modal response of 46.8% (74) who neither agreed nor disagreed.

Respondents' views on situations when it is appropriate to use norm referencing were examined in AQ22i. The results showed that 50.6 % (80) of respondents either disagreed or strongly disagreed with the statement that EAP students who need to demonstrate a particular skill level ought to be assessed and selected based on their rank compared to other students who completed the same test. Only 17% (27) of respondents agreed with the statement whilst 32.2% (51) found themselves unable to either agree or disagree.

The final item in this section, AQ22j, dealt with the practicalities for EAP teachers of converting results of their EAP tests into IELTS or TOEFL grades. The modal response for this item revealed that 32.2% (51) of respondents agreed with the impracticality of the exercise. In total, 44.4% (70) of respondents either agreed or strongly agreed whilst 25.9% (41) disagreed or strongly disagreed, inferring that they find it practical to convert results of EAP tests into IELTS or TOEFL equivalences. Those respondents who were unable to agree or disagree amounted to 29.7% (47).

4.3.2.3.2 Responses to items AQ23-A-J

Item AQ23 contained 10 sub items A-J which were produced in order to collect views from respondents on the topic of their familiarity with various means of statistical analysis. This was an

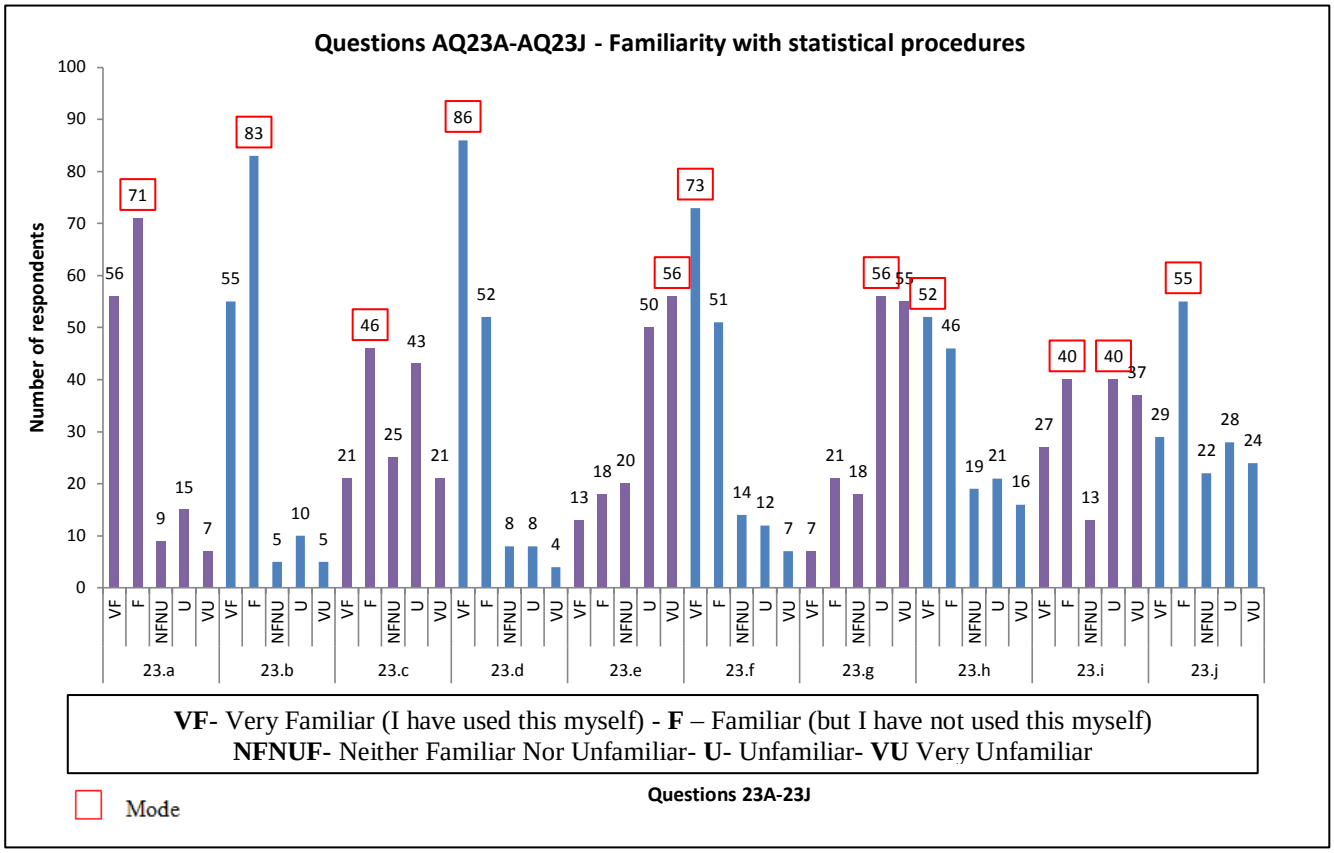
attempt to investigate teachers' usage of more scientific and less subjective approaches to test score analysis (Gorard, 2001, p.5). Table 60 below presents the items which were used:

Table 60 Items/Questions in set AQ23

Item code	Item
AQ23a	How familiar are you with normal distribution
AQ23b	How familiar are you with standard deviation
AQ23c	How familiar are you with Factor Analysis
AQ23d	How familiar are you with Mean
AQ23e	How familiar are you with Cronbach's Alpha
AQ23f	How familiar are you with median
AQ23g	How familiar are you with split half analysis
AQ23h	How familiar are you with mode
AQ23i	How familiar are you with T-tests
AQ23j	How familiar are you with scatter plots

Chart 14 below presents the range of responses provided through items AQ23A-J.

Chart 14 Questions AQ23A-AQ23J for population B



As the questions can be arranged in terms of descriptive or inferential statistics the findings in this section will be grouped accordingly.

With regard to respondent familiarity with descriptive statistics, AQ23a showed that 35.4% (56) of respondents are very familiar with the concept of normal distribution and claim to have made use of the concept in their own practice. The modal response for this item showed that 44.9% (71) of respondents are familiar with normal distribution but have not made use of it in the analysis of their own assessments. 13.9% (22) of respondents were unfamiliar with this form of statistics to some extent, leaving 5.6% (9) who were neither familiar nor unfamiliar.

In AQ23b, respondents were asked about their familiarity with standard deviation, in response to which a similar pattern emerged to AQ23a. 34.8% (55) of respondents noted that they were very familiar with standard deviation and record that they have utilised this statistical procedure during the course of their own testing and assessment. The modal response for this item showed that 53.5% (83) of respondents claim to be familiar with standard deviation but have not yet made use of it in the interpretation of their own assessments. 9.4% (15) of respondents report to be unfamiliar with standard deviation, whilst 3.1% (5) are noncommital.

AQ23d inquired about respondent familiarity with the use of mean or average. The results demonstrate that 54.4% (86) of respondents, the modal response, are very familiar with calculating the mean and have used it in their own interpretations of test scores. In total 87.3% (138) respondents declared some form of familiarity with this concept.

When considering respondent familiarity with the use of the median, AQ23f indicates that the modal response is ‘Very Familiar’ with 46.2% (73) of respondents having employed this statistical procedure during their own interpretations of EAP test scores. In total 78.4% (124) respondents reported a level of familiarity with this area of statistics.

In response to AQ23h, the number of respondents who reported to be very familiar with usage of mode in statistics amounted to 32.9 % (52) of the total respondents. This category represented the modal response to this question. When considered in conjunction with those respondents who were familiar with the mode but who had not used the procedure, the total percentage reached 62% (98). This left 23.4% (37) of respondents who were either unfamiliar or very unfamiliar with the use of the mode and 12% (19) who reported that they were neither familiar nor unfamiliar.

With regard to the use of scatter plots, which can be used to plot points and highlight the relationship between two data sets, the distribution of responses in AQ23j was less varied. The modal response showed that 34.8% (55) of respondents were familiar with scatter plots but had not yet used them. Only 18.4% (29) of respondents were very familiar with scatter plots and had actually used them. The remaining categories of 'neither familiar nor unfamiliar', 'unfamiliar' and 'very unfamiliar' showed respondents with 13.9% (22) 17.7% (28) and 15.1% (24) respectively.

Moving on to inferential statistics, AQ23c focused on respondent familiarity with Factor Analysis, which allows the explanation of relationships between observed variables in terms of a smaller number of unobserved variables (Bachman, 2004, p.111-112). The findings demonstrate that only 13.2% (21) of respondents are very familiar with the procedure and have used it themselves. Although the modal response shows 29.1% (46) of respondents are familiar with the procedure, these individuals do not claim to have actually used the procedure. A total of 40.5 % (64) record that they are either unfamiliar or very unfamiliar with Factor Analysis. The remaining 15.8% (25) of respondents were neither familiar nor unfamiliar with this statistical technique.

In relation to Cronbach's Alpha, which can be used to analyse the internal reliability of a set of items (Dörnyei, 2007, p.206), AQ23e shows that 67% (106) of respondents were either unfamiliar or very unfamiliar with the procedure. This left just 8.2 % (13) of respondents claiming to be very familiar with the procedure and having used it in their assessment practice. A further 11.4% (18) of respondents reported to have some familiarity with Cronbach's Alpha but not to have used it.

Item AQ23g asked respondents about their familiarity with split half analyses, A consistency measure which involves splitting a test into two halves and then comparing the scores from each half of the test to each another (Fulcher & Davidson, 2007, p.105). The distribution of responses shows that 70.2% (111) of respondents are either unfamiliar or very unfamiliar with this procedure. The modal response was 'Unfamiliar' with 35.4% (56) of respondents. Only 13.2 % (21) of respondents reported to be very familiar with split-half analyses to such an extent that they had experience of using this procedure.

Finally, use of T-tests, a method used to determine if the means of two groups of test scores are statistically different from one other (Bachman, 2004), was the focus of AQ23i. For this item the distribution of responses demonstrated that 48.7% (77) of respondents were either unfamiliar or very unfamiliar with this procedure. Whilst 25.3% (40) of respondents declared familiarity with T-tests only 17% (27) of respondents claimed to be very familiar and to have used this form of statistical analysis themselves. The data demonstrated a bimodal frequency of response in both the 'familiar, but not used' and 'unfamiliar' categories, each of which amounted to 25.3% (40).

4.3.2.4 Responses to items linked to consideration of ethics in EAP testing and assessment

4.3.2.4.1 Responses to items AQ25-A-J

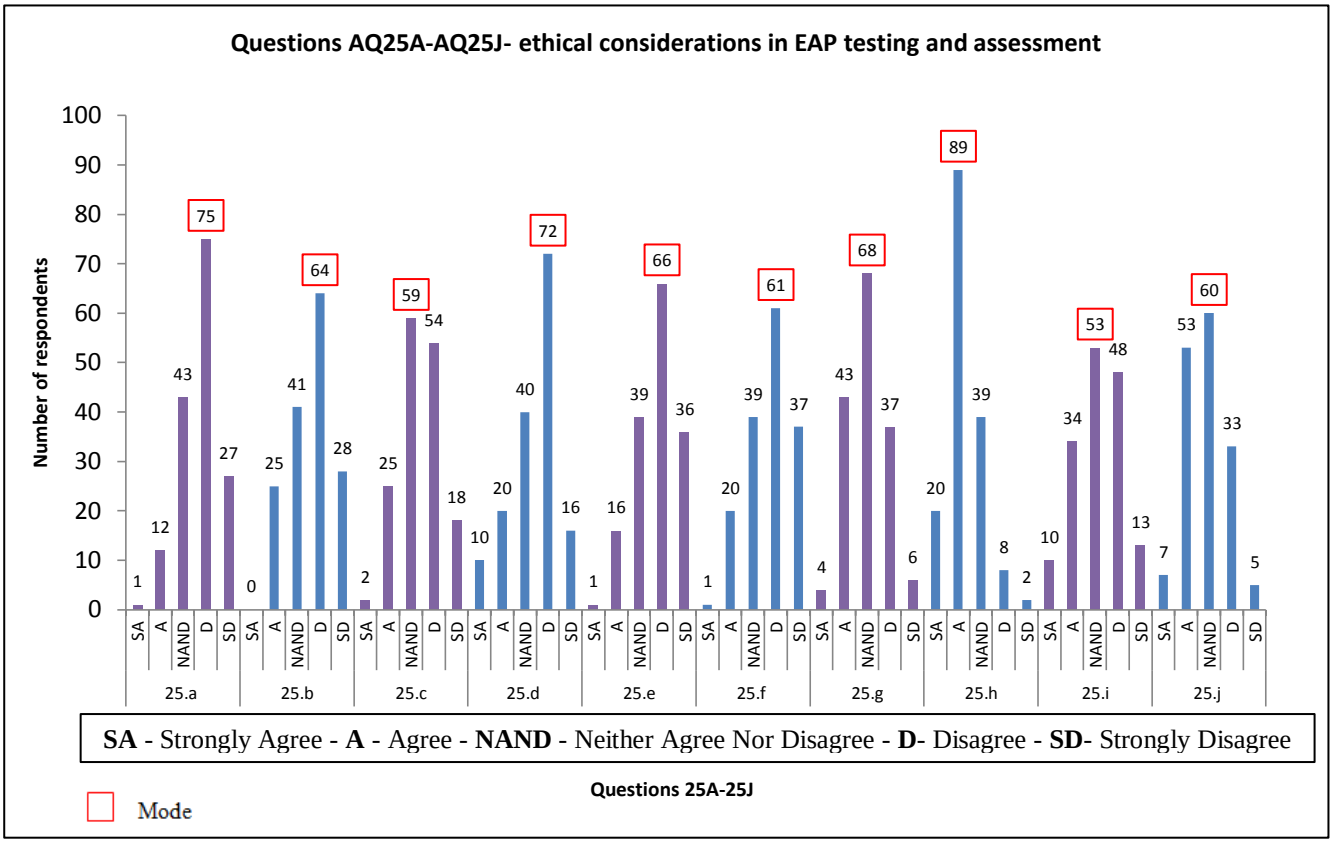
Item AQ25 contained 10 sub items A-J which were created in order to ask participants to comment on their attitudes towards a series of ethical dilemmas. These situations described a series of EAP-assessment-related scenarios which can affect test-takers' lives in potentially adverse ways, such as those highlighted by Shohamy, (2001). Table 61 below lists the items which were included:

Table 61 Items/Questions in set AQ25

Item code	Item
AQ25a	Chinese students struggle with EAP speaking tests and should be required to gain higher grades in EAP speaking tests from next year.
AQ25b	The University has asked to use the Pre-sessional EAP test with a group of seeking to find employment at a local hospital. The Academic Language Centre should allow this.
AQ25c	A security risk has required the evacuation of the building for 15 minutes during group A's EAP test, as a result the tests should be marked as they were at the point of evacuation.
AQ25d	Two tutors from the EAP team would like to give extra classes to certain students who they feel need to improve their examination technique. This action should be avoided.
AQ25e	A University department wants to recruit fewer international students, so the Academic Language Centre should permit the change of the passing grade for current students from 60 to 70%.
AQ25f	Government policy has changed, so it is acceptable to tell a current class of EAP students that they either need to improve their admissions scores by 10%, or return home before the end of their EAP course.
AQ25g	Unexpectedly, nearly all students fail the final EAP test so the EAP team should consider adjusting the marks using a bell curve.
AQ25h	An EAP tutor is preparing students for an EAP test. It is therefore good practice for the tutor to highlight key skills for revision in the class course textbook.
AQ25i	University rules require students to show their library card at the door of the Exam room. Wen Hu has lost his library card so he should not be allowed to enter the room.
AQ25j	Students should be encouraged to question the grades they get in EAP tests.

Chart 15 below presents the range of responses provided through items AQ25A-J.

Chart 15 Questions AQ25A-AQ25J for Population B



Question AQ25a invites respondents to comment on a situation which could lead to an ethical compromise on the grounds of nationality-related discrimination. 64% (102) of respondents disagreed or strongly disagreed with the suggestion that members of a particular nationality should be required, as a group, to gain higher grades in EAP tests. The modal response was 47.4% (75) in the 'Disagree' category. Notably, only 8.2% of (13) respondents either agreed or strongly agreed with this suggestion. 27.2% (43) of respondents were unable to agree or disagree on the best solution.

The dilemma presented in AQ25b is related to decisions associated with test purpose in EAP (Alexander et al., 2008) and surrounding social implications (McNamara & Roever, 2006). When asked if a group of nurses seeking to find employment should be allowed to use a pre-sessional

EAP test, 58.2% (92) of respondents either disagreed or strongly disagreed. The modal response wad attributed to 'Disagree' with 40.5% (64). Only 15.8% (25) of respondents agreed with this suggestion, leaving no respondents who strongly agreed and 25.9% (41) who were unable to either agree or disagree.

Issues relating to fairness and reliability were introduced via AQ25c, which asked respondents to consider whether tests should be marked as they were at the point of an early termination, as a result of a building evacuation for security purposes. Results to this question showed that 45.6% (72) of respondents either disagreed or strongly disagreed that the tests should be marked as they were at the point of evacuation. Only 17% (27) either agreed or strongly agreed. The modal individual response showed that 37.3 % (59) were unsure as to the best manner in which to proceed.

AQ25d was designed to determine respondents' views on fairness with regard to test preparation and examination technique. This item highlighted the dangers of partiality, if some students are made more familiar with the testing format or procedure than others, in advance of the test. Notably, the modal response was 'Disagree' with 45.5% (72) of respondents disagreeing with the need to avoid identifying certain students to whom extra classes should be given on exam technique. In total 52.5% (83) of respondents disagreed with the need to avoid this intervention whilst only 19% (30) of respondents either agreed or strongly agreed.

The situation described in AQ25e outlines a difficult series of circumstances which could arise if the criteria for progression from EAP study to academic study at a higher level are changed during the course of a current EAP cohort's tuition. For instance, this could present an ethical problem if admissions decisions onto the EAP course have been made on the basis of students' need to attain

a lower exit proficiency threshold which has subsequently been raised. 64.6% (102) of respondents either disagreed or strongly disagreed with the raising of the passing grade for current students with the modal response lying with 'Disagree' at 41.7% (66). Only 10.8% (17) of respondents either agreed or strongly agreed with the motion to raise the passing grade.

The quandary described in AQ25f relates to a change in government policy which affects the appropriateness of grades used for admission purposes on EAP courses. Results show that 62% (98) of respondents disagree that a change in government policy should require students to demonstrate a rise in their entry proficiency or return home before the end of their EAP course. Only 13.2 % (21) of responses agreed or strongly agreed that this course of action would be ethically appropriate and 24.6% (39) were unable to decide.

In AQ25g, respondents were asked to reflect on a situation where nearly all students fail an EAP test. Respondents were asked to contemplate if they would consider adjusting the set of class marks with the use of a bell curve, as a means of adjustment, according to a model of normal distribution. 29.7% (47) of respondents either agreed or strongly agreed that using the bell curve to adjust marks would be a suitable course of action whereas, 27.2% (43) of respondents disagreed or strongly disagreed. Notably the most frequent response was 'Neither Agree Nor Disagree' which was selected by 43% (68) of respondents.

With the preparation of students for tests in mind, AQ25h asks respondents if it is good practice for EAP teachers to highlight key skills for revision. 68.9% (109) of respondents indicated agreement or strong agreement with this action, the modal response being 'Agree' at 56% (89). Just 6.3% (10) of respondents expressed either disagreement or strong disagreement. 24.6% (39) of respondents were unable to either agree or disagree.

Respondents are asked, in AQ25i, to consider a situation whereby a test-taker could be excluded from a test session based on failure to provide a library card as proof of identity and contravention of an institutional regulation. This question resonates with what Foucault (1977) and Shohamy (2001) refer to as ceremonial power. 27.8% (44) of respondents agree or strongly disagree that the test taker should be prevented from taking the test, whilst 38.6% (61) either disagree or strongly disagree. The modal individual response shows that 33.5 % (53) are unsure as to how to proceed.

The final ethically-related item in this section, AQ25j, poses the question whether respondents feel that students should be encouraged to question the grades they obtain in EAP tests, part of Assessment Literacy which encourages the empowerment of test takers (Price et al., 2012). Whilst 37.9% (60) of respondents either strongly agreed or agreed with this suggestion, an equal number of 37.9% (60) of respondents selected the modal response of 'Neither Agree Nor Disagree'. 24% (38) of respondents disagreed or strongly disagreed.

4.3.2.4.2 Responses to items AQ26-A-J

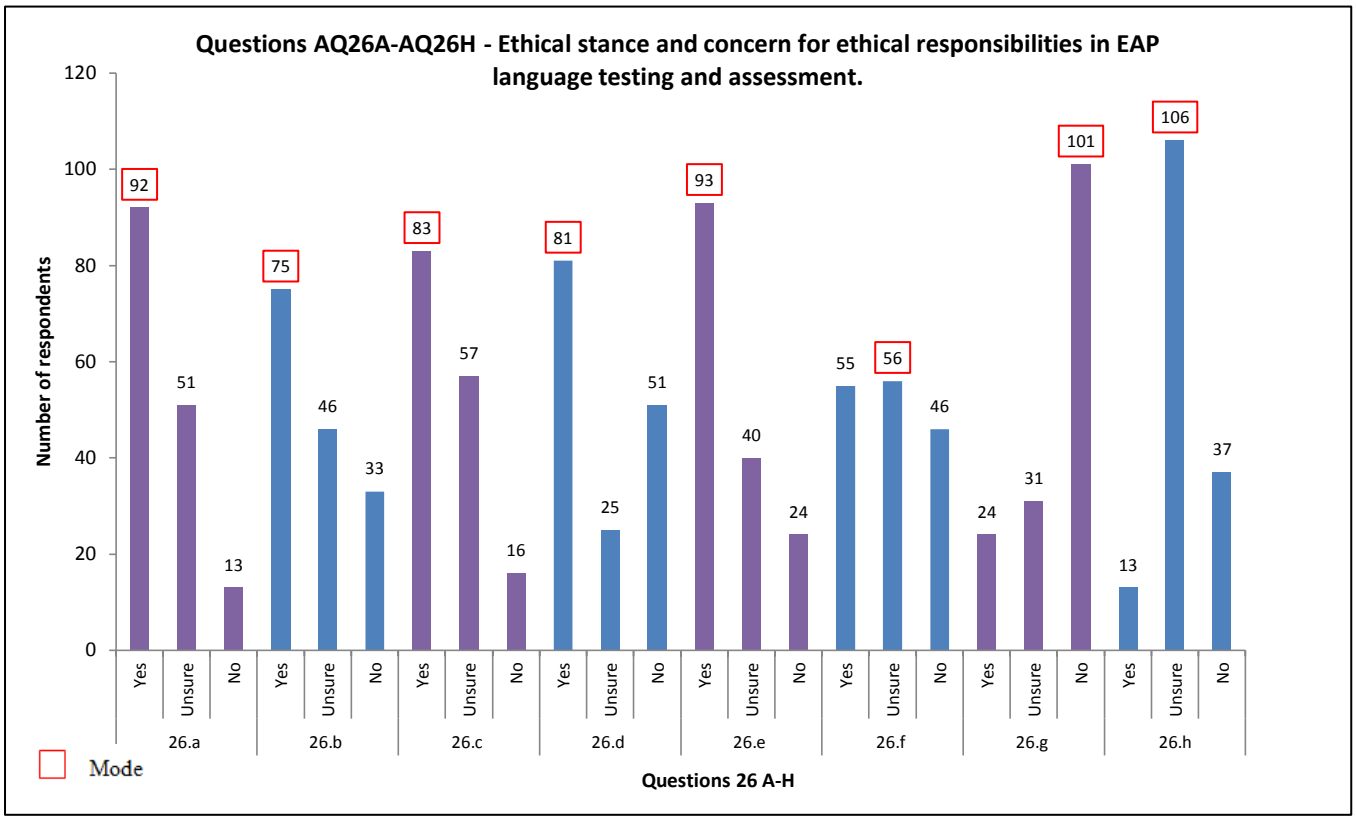
Item AQ26 contained 10 sub items A-H which were created in order to ask respondents to comment more personally with regard to their experience of ethics in practice in EAP assessment. Table 62 below presents the questions which were provided.

Table 62 Items/Questions in set AQ26

Item code	Item
AQ26a	I am familiar with codes of practice for language testing.
AQ26b	I am concerned about the social implications of EAP language testing.
AQ26c	Language testers should make it clear that they have all the relevant knowledge regarding language assessment.
AQ26d	I am responsible for evaluating the ethical consequences of the projects I'm involved with.
AQ26e	I feel able to refuse to participate in procedures which would violate ethical principles of language testing.
AQ26f	I feel that my employers are sufficiently aware of ethical dilemmas associated with language testing.
AQ26g	I have recently felt obliged to contribute to EAP testing procedures which I question the ethicality of.
AQ26h	New regulations connected to language qualification and visa issuance have been introduced in an ethical manner.

Chart 16 below presents the range of responses provided through items AQ26a-h.

Chart 16 Questions AQ26A-AQ26H for population B



AQ26a seeks to gauge respondents' familiarity with codes of practice associated with language testing (ILTA, 2000). 58.2% (92) of respondents claimed to be familiar with language testing codes of practice whilst 32.2% (51) of respondents were unsure with regard to their level of familiarity. This left 8.2% of (13) respondents who were unfamiliar with these codes.

With respect to the social implications of language testing (McNamara & Roever, 2006), 47.4% (75) of respondents to AQ26b declared that they were concerned whilst 20.8 % (33) were unconcerned and 29.1% (46) were unsure regarding their level of concern.

Item AQ26c aimed to investigate the extent to which respondents feel that language testers hold all the relevant knowledge relating to language assessment, a question which may also relate to

engagement with academic disciplines in EAP test development (Fulcher and Davidson, 2007, p.86). The responses to this question show that 53.5% (83) of respondents agree with this statement. Only 10.1% (16) of respondents disagreed with this statement but 36% (57) were unsure.

EAP tutor personal responsibility for evaluating the ethical consequences of testing projects was the focus of AQ26d. The frequency of responses showed that 51.5% (81) of respondents agreed that they were personally responsible whilst 32.4% (51) did not accept that they were responsible. In total, 48.4% (76) were unable to agree that they were responsible for evaluating the ethical consequences of testing projects.

When asked, in AQ26e, if they felt able to refuse to participate in procedures which would violate the ethical principles of language testing, 59.2% (93) agreed and whilst only 15.2% (24) of respondents did not feel able to refuse a further 25.4% (40) of respondents remained unsure.
Responses to item AQ26f, which asked respondents if their employers were sufficiently aware of ethical dilemmas associated with language testing, showed a more uniform spread of responses across each of the three options. A near bimodal response was revealed in both the 'Yes' and 'Unsure' categories which were selected by 35% (55) and 35.6% (56) of respondents respectively. 'No' was selected by 46 respondents, amounting to 29% of the total population.

The frequency of responses to AQ26h demonstrates that whilst 64% (101) denied feeling obliged to contribute to EAP testing procedures, which they question the ethicality of, 15.2% (24) of respondents selected the 'No' option and 19.7% (31) remained unsure. Collectively this suggests that 34.9% of respondents felt unable to fully deny feeling obliged to contribute to EAP testing procedures of questionable ethicality.

The final question in this series was designed primarily to focus on recent changes linked to language testing and visa issuance in the UK. Consequently, the practicality of this question may be questionable for respondents based outside the UK. When asked if new regulations connected to language qualification and visa issuance have been introduced in an ethical manner, 67.5% (106) of respondents selected the 'Unsure' option and 23.5% (37) answered by selecting 'No'. Only 8.2% (13) of respondents were able to agree that new regulations connected to language testing and the issue of visas was introduced ethically.

4.3.2.5 Responses to items linked to influence of research and other resources for the purpose of EAP testing and assessment on EAP teacher assessment practices

4.3.2.5.1 Responses to item AQ13-D

Although AQ13d is a lone item it can be considered to relate to the impact of research on the EAP assessment practices of respondents. The item invites respondents to comment on the regularity of their reference to books or articles associated with language testing and assessment. This is an important aspect of Assessment Literacy which involves taking steps to build expertise and understand more fully what and how to assess (Stiggins, 1995). Only just over one third (34.8%) of respondents agreed or strongly agreed that they consulted such resources with regularity.

4.3.3 Section conclusion

In this section of the Findings chapter, patterns of response to individual items have been revealed, allowing a detailed presentation of respondent views on subconstructs of wider aspects of EAP assessment, which can be related to Assessment Literacy. When it comes to analysis in Chapter 5, features of such responses can add detail for further support or consideration when the analysis of

clusters of interrelated responses, that represent broader constructs within Assessment Literacy, have been taken into account. It is the clustered or composite results which will be described in Section 4.4.

4.4 Set A3- Composite results of questionnaire Likert scales and associated statistics

4.4.1 Section introduction

The questionnaire used in this research has been compiled with a view to collecting information about EAP teacher Assessment Literacy, according to certain sub-constructs. Consequently, the results of responses to the questionnaire are of particular interest in terms of clusters of questions which investigate a common factor or trait, in addition to patterns of response which show commonality in terms of respondent behaviour. As a result this section will focus on the composite results drawn from Likert scales in the questionnaire. Before the composite results are presented a process of improving the validity and reliability of the data which is considered in the thesis will be explored through Factor Analysis and Cronbach's Alpha.

4.4.2 Detailed example of use of Factor Analysis and Cronbach's Alpha using AQ13

According to Dörnyei (2003, p.11) a number of tools exist to assist the researcher in making muti-item scales homogenous.These mechanisms include Factor Analysis, in order to measure construct validity and Cronbach's Alpha reliability coefficients, for the purposes of identifying internal consistency. The key distinction is made by Cortina (1993) who explains that a set of items may be interrelated and multidimensional. It is therefore useful to consider the results of both Factor Analysis and Cronbach's Alpha as Factor Analysis can identify unidimesionality in terms of construct and Cronbach's Alpha identifies interrelatedness of questionaire response to particular items and test length. Consequently, it was decided that results of Factor Analysis and Cronbach's Alpha should be employed in conjunction with researcher intuitive contention (Dörnyei, 2003,

p.111). It was hoped that this would serve to identify groups of items, the results of which can be seen to demonstrate a form of homogeneity and which can be considered in terms of composite multi-item scales.

Individual Factor Analyses and Cronbach's Alpha calculations were conducted for each set of sub items pertaining to DVAS items AQ13, AQ14, AQ16, AQ19, AQ20, AQ22, AQ23, AQ25 and AQ26. SPSS was used for this purpose. The full results of each of these calculations are available in the appendix of this document in section 8.7 on page 449.

Once the results of statistical analysis were computed, the researcher then evaluated the data from each procedure in order to consider which items could be considered sufficiently homogenous to be grouped into a multi item scale.

Brown (2001, p.184) explains that, in the process of analysing a survey instrument for validity, Factor Analysis can be used to explore the convergent and discriminate structures of subsections within survey instruments. With regard to results obtained via Factor Analysis, the researcher analysed the items in each distinct bank of questions which had been grouped with the intention of representing a series of different constructs. The process enabled a more scientific identification of which sub-items did actually group to form factors or constructs and which items did not group as the researcher expected.

In order to identify which Likert scale items clustered into common factors or construct areas, and to identify those items returning sufficiently high eigenvalues according the factors which were discerned, the Rotated Component Matrix output from SPSS was consulted for each series of

items. The matrix from AQ13 and its items A-J, focusing on teachers' skill development and interest and confidence in Assessment Literacy, is provided in Table 63 below as an example.

AQ13 relates to the section of the questionnaire which focuses on training, skills, strengths and weaknesess in EAP testing and assessment. The results of Factor Analyis and Cronbach's Alpha for AQ14, which is also connected to the same aspect of Assessment Literacy are presented later in Section 4.4.3.1.

Table 63 Rotated Component Matrix –Factor Analysis Output for AQ13A-J

Rotated Component Matrix[a]

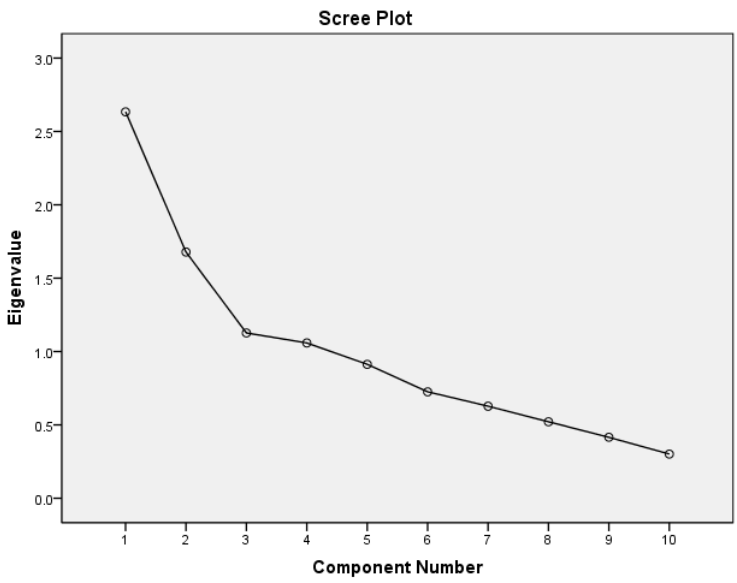
Items in AQ13	Component			
	1	2	3	4
Q13_a	.070	-.043	-.027	.895
Q13_b	.125	.055	.728	.145
Q13_c	.159	-.164	.741	-.188
Q13_d	.624	-.034	.288	.226
Q13_e	.645	.018	.366	.451
Q13_f	.790	-.054	.046	-.201
Q13_g	.699	-.075	.016	.098
Q13_h	-.420	.487	.409	.169
Q13_i	-.050	.866	.026	-.092
Q13_j	-.015	.800	-.140	.014

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalisation.
a. Rotation converged in 6 iterations.

The main columns, numbered 1-4 in Table 63 denote the main factors or components which have been identified within the set of items. The numbers set against each question item in each of the four columns represent eigenvalues which can be used to determine the extent to which each item loads on the factors or components identified. It is often considered that a variable makes a significant contribution to a factor if the loading is 0.3 or greater (Hinton, 2004, p.341). The factors in which each of items AQ13A-J load most highly and above 0.3 have been highlighted in green, in Table 63, in order to identify meaningful factor-based clusters.

A second step in the process of Factor Analysis involves the consultation of a Scree Plot which is also produced by SPSS. The Scree Plot for AQ13A-J is presented below in Chart 17 below.

Chart 17 Scree plot from Factor Analysis of AQ13A-J



According to Dörnyei (2007, p.235) only the factors on the slope of the Scree Plot should be considered. In this case this would indicate the first three factors.

As a result, this process has revealed three potential factors within AQ13. Factor 1 is comprised of items D,E,F and G. Factor 2 contains items H, I and J and Factor 3 includes items B and C.

When the nature of the actual items is considered further, it is clear that Factor 1 relates to 'Respondent engagement and confidence in EAP language testing and assessment', whilst Factor 2 can be interpreted as 'Respondents' perceptions regarding complexity of EAP language testing and assessment'. Notably Factor 3 is less easy to understand in terms of a single construct.

Table 64 below shows factor groupings and the name of each factor against each of the original items.

Table 64 Factors and factor names identified for AQ13

Item code	Item	Factor	Factor Name
AQ13b	Testing and assessment skills are crucial for my role as an EAP tutor.	Factor 3	?
AQ13c	The skills that I have acquired in EAP assessment have mainly been developed through courses I have taken which include focus on language testing.		
AQ13d	I regularly read books or articles on language testing.	Factor 1	Respondents' interest and confidence in EAP language testing and assessment
AQ13e	EAP testing and assessment is one of my key skill areas.		
AQ13f	The reference and guidance material available to me on the subject of language testing and assessment is very user-friendly.		
AQ13g	I feel that I have had sufficient training in EAP assessment and testing practices.		
AQ13h	It is more straightforward to test skills in EAP than to test subject knowledge.	Factor 2	Respondents' perceptions regarding complexity of EAP language testing and assessment
AQ13i	It is straightforward to identify the key EAP language areas and skills that students need to be assessed in.		
AQ13j	It is straightforward to recognise a question which is not working properly in an EAP test or assessment.		

According to Brown (2001, p.176), for survey instruments which have distinct subsections, examining the reliability of each of the subsections is much more important than examining the reliability of the survey as a whole, given that different sections are designed to measure different constructs.

Once the Factor Analysis for this set of items was completed, Cronbach's Alpha was calculated in SPSS on each of the three factors identified from AQ13, using only those items which were identified as loading most highly within them. It was noted that any items which were phrased negatively should be reversed at this stage. For AQ13, this included items A, H, I and J from AQ13.

The output from SPSS relating to Cronbach's Alpha for AQ13 in Factors 1, 2 and 3 can be viewed in Figure 24 below.

Figure 24 Cronbach's Alpha reliability statistics for Factors 1,2& 3 from Item set AQ13

Reliability Statistics For AQ13 Factor 1

Cronbach's Alpha	Cronbach's Alpha Based on Standardised Items	N of Items
.720	.722	4

Item-Total Statistics For AQ13 Factor 1

Item-Total Statistics For Q13 factor 1

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q13_d	9.05	5.310	.500	.348	.665
Q13_e	9.27	5.027	.637	.437	.576
Q13_f	9.13	6.447	.455	.215	.691
Q13_g	9.13	5.606	.460	.265	.689

Reliability Statistics For AQ13 Factor 2

Cronbach's Alpha	Cronbach's Alpha Based on Standardised Items	N of Items
.614	.613	3

Item-Total Statistics For AQ13 Factor 2

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q13_h	5.98	2.885	.328	.157	.646
Q13_i	5.59	2.256	.571	.331	.281
Q13_j	5.42	2.818	.384	.229	.567

Reliability Statistics For AQ13 Factor 3

Cronbach's Alpha	Cronbach's Alpha Based on Standardised Items	N of Items
.420	.446	2

Item-Total Statistics For AQ13 factor 3

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q13_b	3.13	1.377	.287	.082	.
Q13_c	1.92	.624	.287	.082	.

In this example, it can be seen from the Item-Total Statistics tables, which relate to each of the three factors that Cronbach's Alpha for Factors 1, 2 and 3 amount to .720, .614 and .420 respectively. When the Item total statistics tables are consulted for each factor it is evident for factors 1 and 2 that removing any of the items from within the factor would not increase the alpha. Accordingly, as only two items are comprised within Factor 3, neither of the items can be deleted as alpha cannot be calculated on a lone item.

When considering the alpha levels which will be judged as acceptable in this study, it has been determined that whilst lower levels are common on shorter scales which are usual in questionnaires which seek to measure many different areas, concern should be raised in situations where alpha falls below 0.6 (Dörnyei, 2003, p.112). Nevertheless, with reference to advice provided Tuckman (1999, p.445) a Cronbach's Alpha of 0.5 or above may be considered acceptable for scales which focus on attitude or preference assessments.

As a result of the information revealed in these findings, it was decided that due to the difficulty in associating Factor 3 from AQ13 with any meaningful unidimensional construct and given the lower Cronbach's Alpha for these items, they should be withdrawn from the study.

Consequently, this leaves Factors 1 and 2 which will be used to calculate composite scales results for each of the items within each factor. The results of the remaining Factor Analyses and Cronbach's Alpha calculations are summarised in the following sections of this chapter.

Following the same process as described in this section, the remaining groups of items were analysed using Factor Analysis, Cronbach's Alpha and the researcher's intuitive contention.

4.4.3 Results of Factor Analysis and Cronbach's Alpha for remaining items linked to training, skills, strengths and weaknesses in EAP testing and assessment design

4.4.3.1 Factor Analysis and Cronbach's Alpha for AQ14

The intended questionnaire section construct for AQ14 was 'Respondent skills in aspects of EAP assessment'; to examine perceptions regarding skills related Assessment Literacy. It should be noted that AQ13 and AQ14 both relate to the section of the questionnaire which investigates training, skills, strengths and weaknesses in EAP testing and assessment. The findings for AQ13 were however described in Section 4.4.2 in order to provide a detailed description of the process followed.

After running a Factor Analysis on items in AQ14, two factors were identified both through the eigenvalues in the Rotated Component Matrix and through inspection of the Scree Plot. Although each Factor loaded highly, Items AQ14a and AQ14f, which loaded highly in Factor 2, also loaded highly in Factor 1. It was also difficult to conceptualise a separate descriptor for each of Factors 1 and 2 with items AQ14a and AQ14f remaining in factor 2. Consequently, all items were combined into Factor 1 with the exception of items AQ14g and AQ14h which only loaded sufficiently highly on Factor 2. This then also revealed that items AQ14g and AQ14h, now the sole contributors to Factor 2, could allow the factor to be renamed 'Respondent skills in using statistics', thus leaving Factor 1 as 'Respondent skills in aspects of EAP assessment'.

After Cronbach's Alpha was run on Factors 1 and 2, for which no reverse codings were necessary, it was revealed that the alpha for Factor 1 was .922 and the Alpha for Factor 2 was .816.

Consequently, it was considered that both factors 1 and 2 could be used for the purposes of composite scale calculations.

4.4.4 Results of Factor Analysis and Cronbach's Alpha for items linked to approaches to EAP testing and assessment design

4.4.4.1 Factor Analysis and Cronbach's Alpha for AQ16

The overarching section construct which was the intention of AQ16 was 'Respondents' views on EAP test and assessment design processes', to investigate principled approaches to assessment. After reviewing the Rotated Component Matrix for AQ16A-J and consultation of the accompanying Scree Plot, it was ascertained by the researcher that four main factors could be identified. Unlike for AQ14, high cross-loadings across different factors did not identify any noteworthy patterns. After reverse coding Items, C,F and G and conducting Cronbach's Alpha for each of the four factors identified, it was noted that both factors 2 and 3 showed negative covariance.

With regard to Factors 1 and 4, these revealed a Cronbach's Alpha of .538 and .566 respectively, which are somewhat lower than .6, a commonly selected threshold. However, these levels were sufficiently in line with the .5 level of acceptability as cited by Tuckman (1999, p.445), for scales focusing on attitudes or preferences. After consulting the specific items in AQ16, it was decided by the researcher that items A, F and G could logically be grouped as a factor described as 'Respondents' views on processes and procedures in EAP test design' and that items D and E converged on the topic of 'Respondents' views on specificity in EAP testing and assessment', thus reflecting enduring dichotomies as expressed by Hyland (2006). Consequently, despite the lower

Cronbach's Alpha result for Factors 1 and 4, it was decided to retain these for the composite scale analysis based on the nature of the scale, the results of the Factor Analysis and the researcher's intuition. Factors 2 and 3 were discarded.

4.4.5 Results of Factor Analysis and Cronbach's Alpha for items linked to views on/ experience of validity and reliability in EAP testing and assessment

4.4.5.1 Factor Analysis and Cronbach's Alpha for AQ19

The objective of including the questions in section AQ19 was to ask respondents for their views on validity and reliability in EAP testing and assessment. Through completing the Factor Analysis stage it became clear from the Rotated Component Matrix that four initial factors were identified. Subsequently through examining the Scree Plot it was evident that only three factors should be retained for further scrutiny. After calculating Cronbach's Alpha and reversing coding items B, C, D, G, H, for each of the three remaining factors, only Factor 1 returned a satisfactory alpha with a level of .642. On consultation of the data, it appeared that a suitable new label for Factor 1 would be 'Respondents' views on content validity'.

4.4.5.2 Factor Analysis and Cronbach's Alpha for AQ20

The theme for the set of questions in AQ20 was intended to relate to 'Respondents' views on methods of enhancing validity of inferences drawn from EAP testing and assessment.' When running the Factor Analysis statistics for this set of questions, three initial factors were revealed, however with review of the Scree Plot output from SPSS, it appeared that the curve tailed off after two Factors. On running Cronbach's Alpha with C, D, F, G and H reversed on Factors 1 and 2, it transpired that only Factor 2 revealed an Alpha of a suitable level at .623. Consequently It was

decided that only Factor 2 should be retained and that the original label could be kept as, 'Respondents' views on methods of enhancing validity in EAP testing and assessment'.

4.4.6 Results of Factor Analysis and Cronbach's Alpha for items linked to practices involving analysis and interpretation of the results and scores of EAP tests and assessments

4.4.6.1 Factor Analysis and Cronbach's Alpha for AQ22

AQ22 was designed to ask questions relating to 'Respondents' views on analysis and interpretation of EAP tests & assessment results.' After the Rotated Component Matrix was evaluated, four factors were retained; however on examination of the Scree Plot, the flattening of the curve suggested that two factors were a more accurate estimation of existing components. When these two factors were subjected to a Cronbach's Alpha scale analysis, after having reverse coded items A, C, F, H and I, it was demonstrated that Factor 1 had the highest alpha at .588. The items in this scale were then examined by the researcher and it was judged that an appropriate new label would be 'Respondents' views on Criterion/Outcomes based evaluation and feedback.' Whilst the Alpha value was a recognised as low by the researcher, as both the Factor Analysis and the researcher's consideration of the items within the scale showed a logical fit, it was decided that the Factor was worthwhile retaining.

4.4.6.2 Factor Analysis and Cronbach's Alpha for AQ23

The items in AQ23 were included in the survey with a view to examining respondent familiarity with statistical terms and procedures. After a Factor Analysis, both the Rotated Component Matrix

and Scree Plot showed the presence of two key factors. When the items loading on each factor were considered, it was clear to see that these grouped into two sections, Factor 1 emerged as 'Respondent familiarity with descriptive statistics' and Factor 2 related to 'Respondent familiarity with inferential statistics'. After conducting a Cronbach's Alpha analysis on both these scales, it was discovered that Factor 1 has an Alpha of .863 and Factor 2 reached .861. Consequently both Factors were retained.

4.4.7 Results of Factor Analysis and Cronbach's Alpha for items linked to consideration of ethics in EAP testing and assessment

4.4.7.1 Factor Analysis and Cronbach's Alpha for AQ25

The purpose of the items in AQ25 was to investigate 'Respondents' views on ethical dilemmas in EAP testing and assessment.' With reference to the Rotated Component Matrix, the data returned four initial factors with loadings requiring further consideration. The Scree Plot then indicated that a two factor model would be more appropriate. After reversing items A, B, D, E and F, Cronbach's Alpha showed that Factor 1 had an alpha of .617, whilst Factor 2 had a lower alpha of .475. When referring to the items in Factor 1, it appeared logical to name the factor 'Respondents' views on high-stakes EAP assessment dilemmas.'

4.4.7.2 Factor Analysis and Cronbach's Alpha for AQ26

Section AQ26 was devised in order to determine teachers' personal experience of ethical situations associated with EAP in the work place. With reference to the Rotated Component Matrix, it was apparent that three initial factors were returned. The Scree Plot also indicated three

factors. After considering the items in each identified factor and their interconnection, it was decided that item H should be reloaded from Factor 3 to Factor 1 due to a logical fit and a sufficiently high load in Factor 1.

When the Cronbach's Alpha statistics were run on the remaining factors 1 and 2, it became clear that Factor 1 had the highest alpha with .537 and Factor 2 was at .478.

Consequently, Factor 2 was withdrawn. After consideration of the Factor Analysis and additional scrutiny of the content and focus of the items loading in Factor 1, it was decided that Factor 1 could be retained and named 'Ethics in practice in EAP assessment'.

4.4.8 Overall Review of Factor Analysis and Cronbach's Alpha results

Table 65 below summarises the outcome of Factor Analysis and Cronbach's Alpha which have been used to decide which factors or constructs from within Assessment Literacy to focus on in the composite Likert scale results. After the initial factors were identified, it was discovered that Factor 4 from AQ16 and Factor 1 from AQ19 could logically be summed to form a single factor. Once the Cronbach's Alpha for this was completed it also revealed both an acceptable level at .632 and a longer scale. The tables showing the Cronbach's Alpha calculations for this new factor can be viewed in section 8.7.20 of the Appendix. Similarly as Factor 2 from AQ14 contained only two items, each of which dealt with descriptive and inferential statistics respectively, it was decided to add AQ14G into AQ23 Factor 1 and AQ14H into AQ23 Factor 2. Both of these actions retained a high Cronbach's Alpha .870 for the factor relating to descriptive statistics and .863 for inferential statistics. The charts for this calculation can be seen in 8.7.19 of the Appendix.

Whilst the majority of factors show a Cronbach's Alpha of 0.6 or above, three of the factors show an alpha of above 0.5. As these scales relate to attitude and preference it has been judged that this is acceptable, in light of the minimum accepted alpha for attitude and preference scales as suggested by Tuckman (1999, p.445).

As two of the Factors from Table 65 are based on Factors comprised of only two items, it should be noted that the Cronbach's Alpha for these clusters cannot be judged as markedly informative, given that only two items represents a restricted sampling of a much larger domain (Cudeck, 2001, p.55), nevertheless they have been retained as it is the researcher's belief that they represent sufficiently meaningful constructs which have not been disproven by the Factor Analysis or Cronbach's Alpha calculations.

Table 65 Summary of Factor Analysis and Cronbach's Alpha results

Factor Name	Factor – from Rotated Component Matrix	Items/variables contributing to Factor	Items requiring reverse coding	CB	Notes
'Respondents' interest and confidence in EAP language testing and assessment'	Factor 1	AQ13 Items D,E,F,G	None reversed	.720	Not Applicable
'Respondents' perceptions regarding complexity of EAP language testing and assessment'	Factor 2	AQ13 Items H,I,J	H I J	.614	Not Applicable
'Respondent skills in aspects of EAP assessment'	Factor 1	AQ14 Items A,B,C,D,E,F,I,J	None reversed	.922	Not Applicable
'Respondent skills in using Statistics'	Factor 2	AQ14 Items G, H	None reversed	.816	Add in to Factors from AQ23 as AQ14G = Descriptive statistics and AQ14H = Inferential statistics
'Respondents' views on processes and procedures in EAP test design'	Factor 1	AQ16 Items A,F,G	F G	.538	Not Applicable
'Respondents' views on specificity in EAP testing and assessment'	Factor 4	AQ16 Items D,E	None reversed	.566	These Factors have subsequently been combined into one four-item factor of 'Respondents' views on aspects of Construct validity' as together they return a CB of .632 (.652 if AQ19e is deleted but as this is a negligible difference retain it)
'Respondents' views on content validity'	Factor 1	AQ19 Items E, F	None reversed	.642	
Respondents' views on methods of enhancing validity in EAP testing and assessment	Factor 2	AQ20 Items A, E	None reversed	.623	Only two items
Respondents' views on Criterion/Outcomes based evaluation and feedback in EAP	Factor 1	AQ22 Items D,E	None reversed	.588	Only two items
Respondent familiarity with descriptive Statistics in EAP assessment	Factor 1	AQ23 Items A,B,D,F,H,J	None reversed	.863	AQ14 G added to this factor. Cronbach's Alpha now .870
Respondent familiarity with Inferential statistics in EAP assessment	Factor 2	AQ23 Items C,G,E,I	None reversed	.861	AQ14H added to this factor. Cronbach's Alpha now .868
Respondents' views on high-stakes EAP assessment dilemmas	Factor 1	AQ25 Items ABEF	ABEF	.617	Whole set negatively phrased so no need to reverse
Ethics in Practice in EAP Assessment	Factor 1	AQ26 Items E, F,G,H	G	.537	Not Applicable

4.4.9 Composite Likert Scale Results

The questionnaire tool designed for this research project has been created for the purpose of collecting data related to EAP teacher Assessment Literacy and the extent to which teachers' views reflect research and good practice in language testing. With this in mind, the survey has been organised into a series of measures of sub-constructs of Assessment Literacy. As mentioned in Section 4.4.10, the questionnaire findings are of greater interest and significance when evaluated in terms of groupings of related questions, rather than individual items. This is due to the fact that '…multi-item scales maximize the stable component that items share and reduce the extraneous influences unique to the individual items.'(Dörnyei, 2003, p.34). Consequently, the section which follows will present composite results of related Likert Scale DVAS questions.

4.4.10 Composite Likert Scale results for items linked to training, skills, strengths and weaknesses in EAP testing and assessment

4.4.10.1 AQ13 Factor 1 (DEFG)-'Respondent engagement and confidence in EAP language testing and assessment'

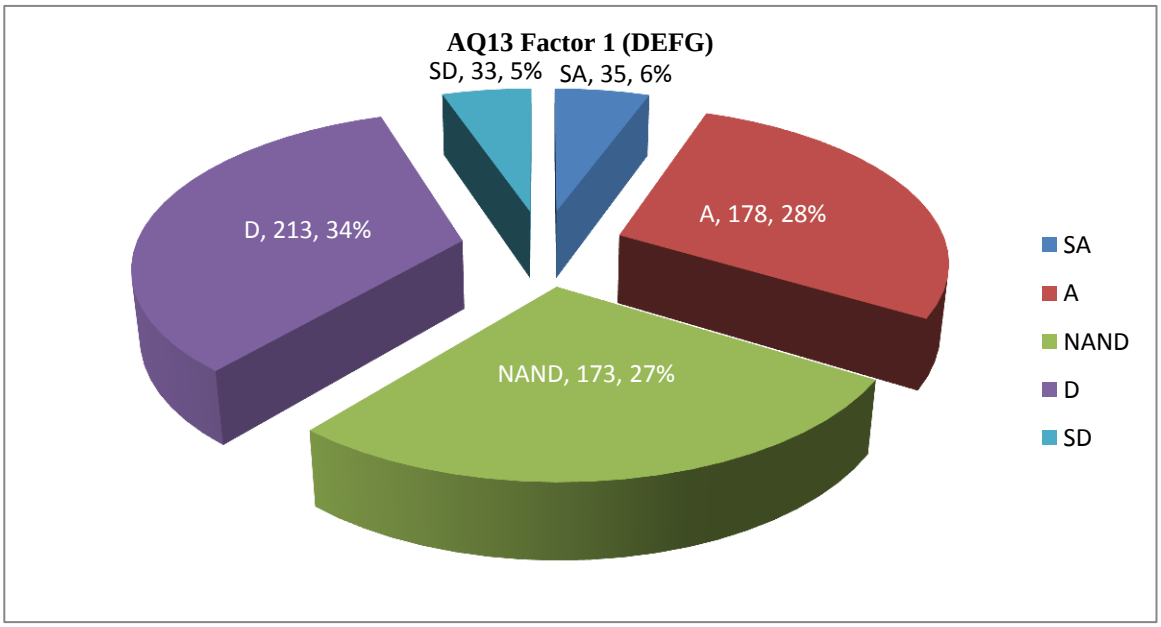
The composite results for Factor 1 within AQ13, comprising items D, E, F, and G deal with questions associated with EAP teachers' engagement with and confidence in EAP assessment. In particular the questionnaire items focus on the regularity with which teachers consult books on language testing, the extent to which respondents feel EAP assessment represents a key skill for them, whether they find support material user friendly and perceptions regarding the need for additional training. The summed results of these Likert items show that only 33.7% strongly agree or agree with items grouped in this category, whilst 38.9% of respondents either disagree or strongly disagree. Although there is almost a multimodal pattern with responses to 'Agree',

‘Neither agree nor disagree’ and ‘Disagree’ located close to 30%, it is notable that only just over one third of respondents were able to agree or strongly agree and thus clearly indicate confidence and engagement with EAP testing and assessment.

Table 66 Composite results for AQ13 Factor 1 Items DEFG

Category of response	Summed Frequency of Response	Percent	Cumulative Percent
SA	35	5.5	5.5
A	178	28.2	33.7
NAND	173	27.4	61.1
D	213	33.7	94.8
SD	33	5.2	100.0
Total	632	100	

Chart 18 Pie chart showing distribution of response to AQ13 Factor 1 (DEFG)



4.4.10.2 AQ13 Factor 2 (HIJ)-'Respondents' perceptions regarding complexity of EAP language testing and assessment

Combined results for Factor 2, taken from AQ13 included items which ask respondents to comment on whether they find aspects of EAP assessment straightforward. The items in this set have been reversed due to the stance adopted by the researcher which purports that the processes described are inherently complex given the multidimensional nature of language assessment. This stand point is supported by Inbar-Lourie (2008), who affirms, with reference to courses in language testing, that account needs to be taken of the organic and complex nature of language testing and its varied contexts, otherwise language assessment risks being rendered ineffective.

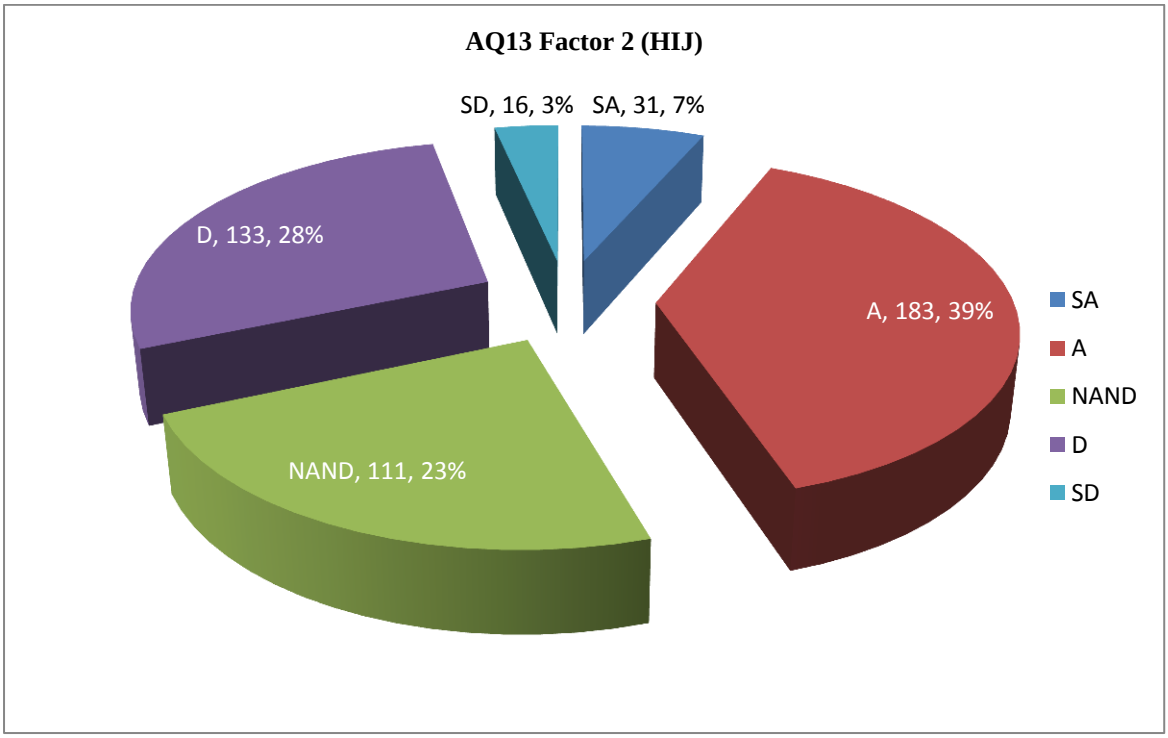
Consequently, this factor can now be considered in terms of whether respondents consider aspects of EAP assessment to be complex, including the identification of the features of EAP in which students need to be assessed, the recognition of faulty test items and the relative complexity of language assessment compared to subject knowledge.

The results show that 45.1% of respondents agree or strongly agree that EAP assessment is complex in the manners described above, whilst 31.5% disagree or strongly disagree. This has left 23.4% of respondents who neither agreed nor disagreed that the identified aspects of EAP assessment were complex.

Table 67 Composite results for AQ13 Factor 2 Items HIJ

Category of response	Summed Frequency of Response	Percent	Cumulative Percent
SA	31	6.5	6.5
A	183	38.6	45.1
NAND	111	23.4	68.6
D	133	28.1	96.6
SD	16	3.4	100.0
Total	158	100	

Chart 19 Pie chart showing distribution of response to AQ13 Factor 2 (HIJ)



4.4.10.3 AQ14 Factor 1 (ABCDEFIJ)-'Respondent skills in aspects of EAP assessment'

The focus of Factor 1 from AQ14 is respondent skills in aspects of EAP assessment, as inspired by the Test Design Framework (Fulcher, 2010, p.94). Individual items in AQ14 represent sub skills within EAP testing and assessment and thus a composite of these results can be used to consider respondents' perceptions regarding their overall skill level in EAP assessment, although it should not be considered that the breadth of skill coverage is comprehensive.

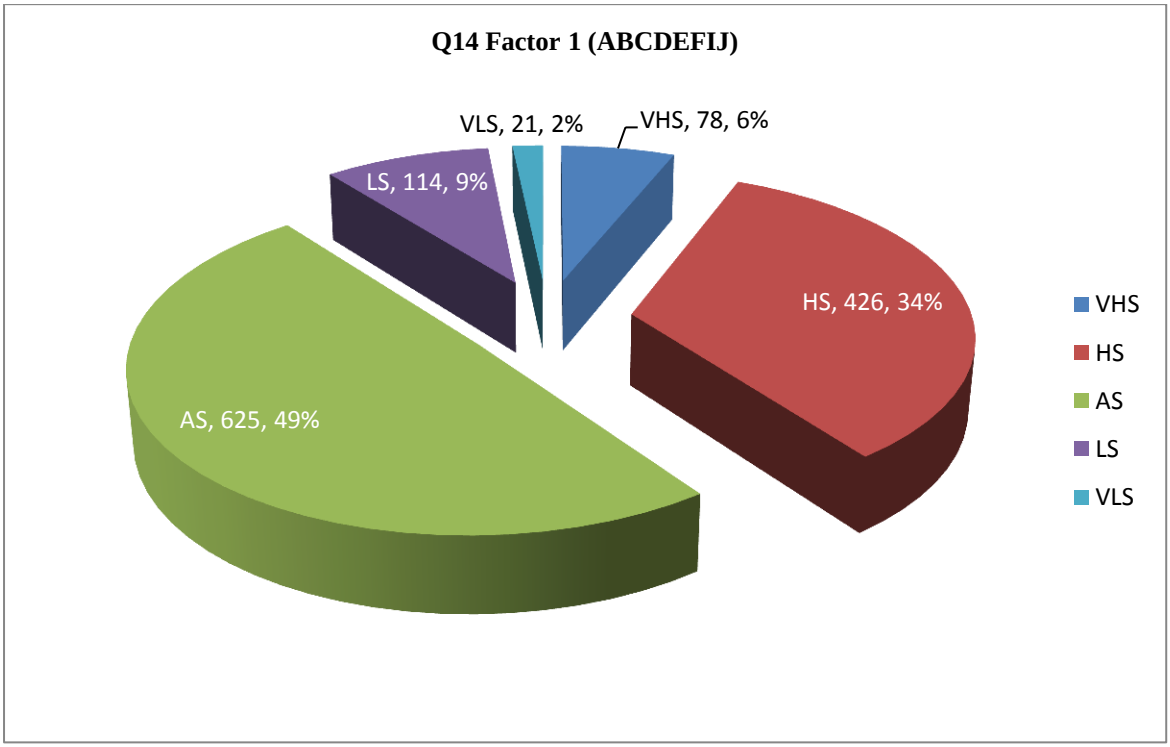
The results of the composites for these items show that 39.8% of respondents consider themselves to be either 'Highly Skilled' or 'Very Highly Skilled'. Although only 10.6% describe themselves as 'Low Skilled' or 'Very Low Skilled', 49.4% rate themselves as being of average skill level in the areas identified.

Further details relating to specific skill areas can be gleaned through consultation of results relating to individual items as presented in Findings Set A2.

Table 68 Composite results for AQ14 Factor 1 Items ABCDEFIJ

Category	Summed Frequency of response	Percent	Cumulative Percent
VHS	78	6.1	6.1
HS	426	33.7	39.8
AS	625	49.4	89.2
LS	114	9.0	98.2
VLS	21	1.6	100
Total	1264	100	

Chart 20 Pie chart showing distribution of response to AQ11 Factor 1 (ABCDEFIJ)



4.4.11 Composite Likert Scale results for items linked to approaches to EAP testing and assessment design

4.4.11.1 AQ16 Factor 1 (AFG)-'Respondents' views on processes and procedures in EAP test design'

Items grouped in this next factor, which are drawn from AQ16, are associated with processes and procedures in EAP test design, in particular the starting point for test development; the need for teamwork (Jafarpur, 2003, p.72), in line with the view of test development as 'an creative, organic, consensus driven, iterative process' (Fulcher and Davidson 2007, p.61) and the beneficial effects of practice and experience over time. Items F and G were reversed in order for the coding of response to align with Item A.

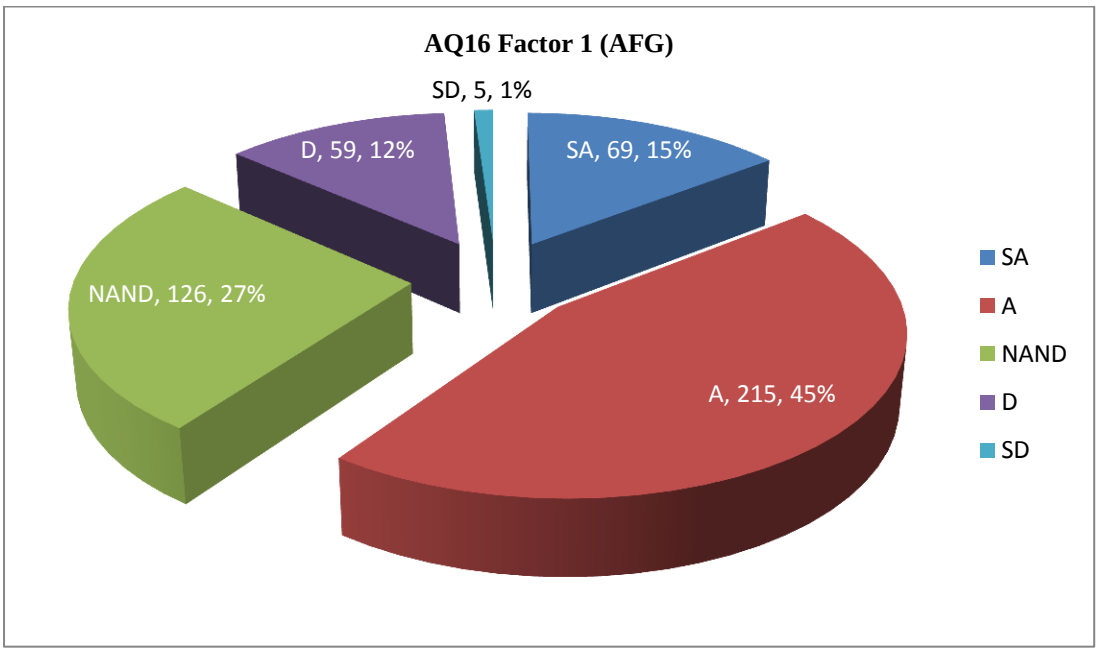
The results demonstrated that 60% of respondents agreed or strongly agreed with statements which indicate an understanding of the importance of starting the test design process at a point other than with writing test items, the significance of teamwork and the benefits to practitioners of gaining experience of test design over time.

Only 13.5 % of practitioners disagreed or strongly disagreed with the significance of the factors mentioned above, leaving 26.6% of respondents who neither agreed nor disagreed.

Table 69 Composite results for AQ16 Factor 1 Items AFG

Category of response	Summed Frequency of Response	Percent	Cumulative Percent
SA	69	14.6	14.6
A	215	45.4	59.9
NAND	126	26.6	86.5
D	59	12.4	98.9
SD	5	1.1	100.0
Total	474	100	

Chart 21 Pie chart showing distribution of response to AQ16 Factor 1 (AFG)



4.4.12 Composite Likert Scale results for items linked to views on/experience of validity and relaiability in EAP testing and assessment design

4.4.12.1 AQ16 Factor 4 (DE) and AQ19 Factor 1 (EF) -'Respondents' views on aspects of construct validity'

Items AQ16 D and E along with Items AQ19 E and F have been grouped together in a category which is deemed to represent 'Respondents' views on aspects of construct validity' with a particular emphasis on the extent to which students studying different subjects and academic fields should have assessments designed to reflect the constructs and idiosyncrasies particular to their domain of study. Seminal articles about this topic have been written by Spack (1988) and Dudley-Evans and St John (1998).

41.2% of respondents either agreed or strongly agreed in response to items in this series, suggesting their support for creating assessments which reflect the constructs common to specific domains of study based on students' area of specialism. 28.9% of respondents either disagreed or

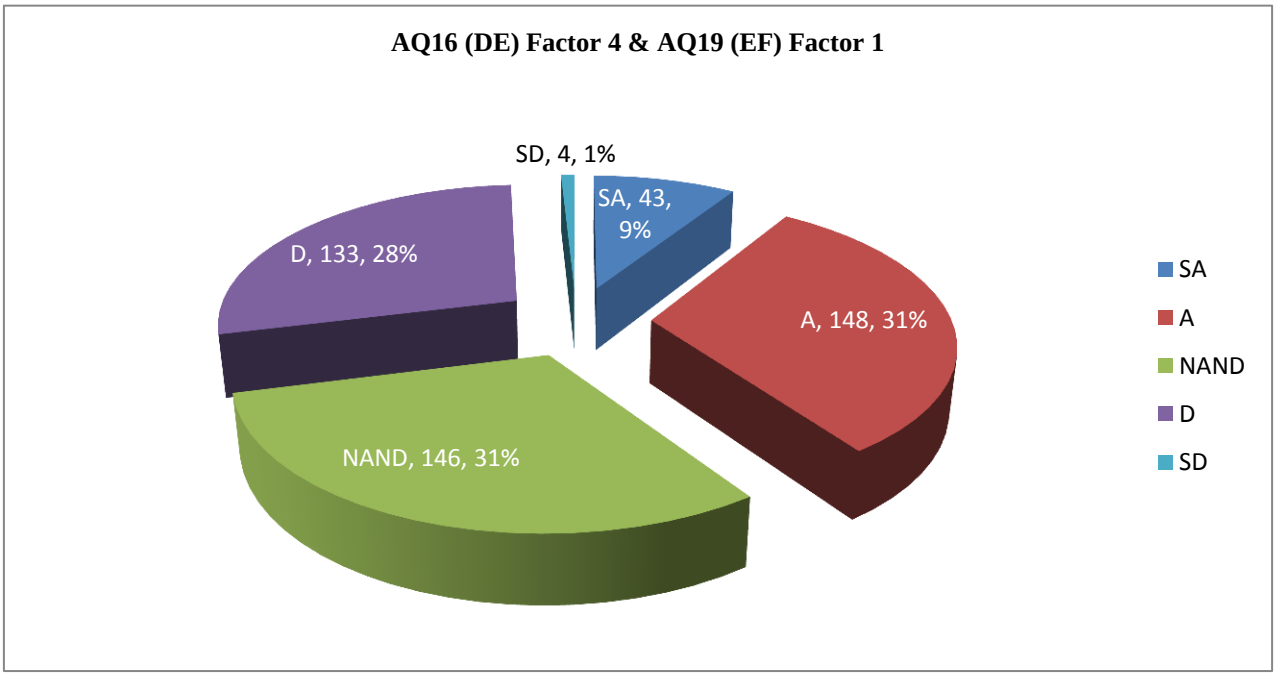
strongly disagreed with the need to maintain construct validity in the manner described. This left 31% who neither agreed nor disagreed.

This range of views could be said to mirror the debate surrounding the choices and challenges of embracing either process or product and as a result, English for General Academic Purposes (EGAP) or English for Specific Academic Purposes (ESAP) (Jordan, 1997; Selinker et al., 1981; Spack, 1988; Hutchinson & Waters in Swales, 1985; Widdowson, 1983).

Table 70 Composite results for AQ16 Factor 4 (DE) and AQ19 Factor 1 (EF)

Category of response	Summed Frequency of Response	Percent	Cumulative Percent
SA	43	9.1	9.1
A	148	31.2	40.3
NAND	146	30.8	71.1
D	133	28.1	99.2
SD	4	0.8	100.0
Total	474	100	

Chart 22 Pie chart showing distribution of response to AQ16 Factor 4 (DE) and AQ19 Factor 1 (EF)



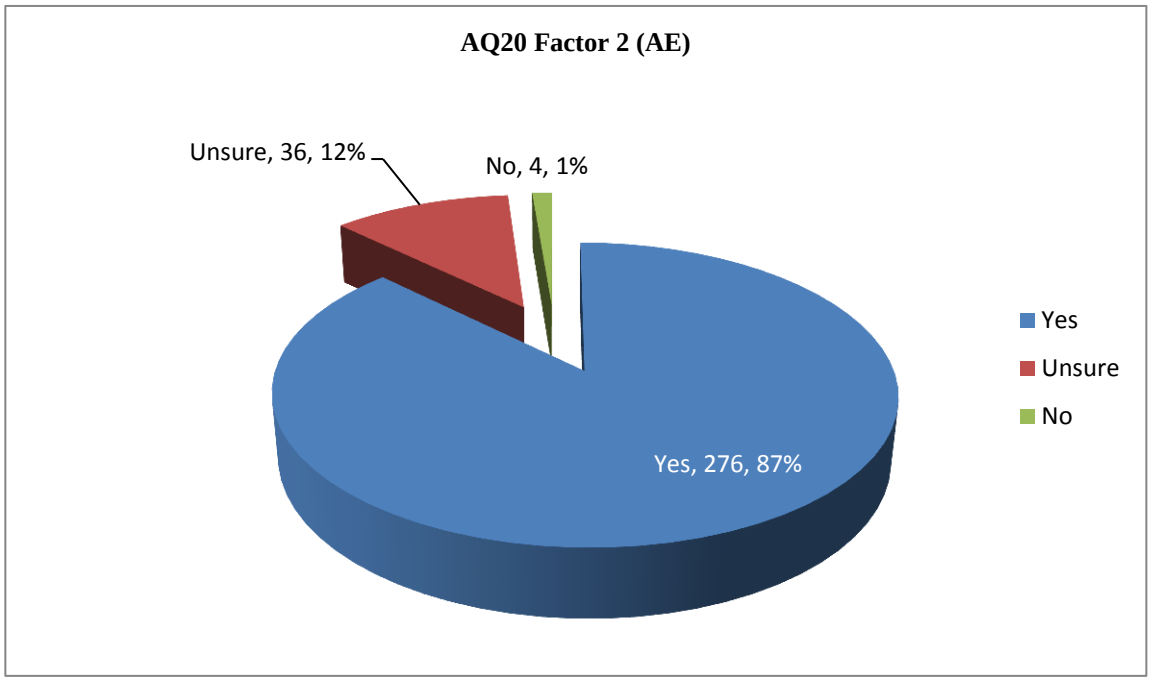
4.4.12.2 AQ20 Factor 2 (AE) - ‘Respondents’ views on methods of enhancing validity in EAP testing and assessment.

The two questions which have been grouped for AQ20 related to methods of enhancing validity of inferences from test results. More specifically, the items ask respondents to comment on the extent to which they agree or disagree that the validity of inferences from tests can be enhanced through research into students’ area of academic study, in addition to collaboration with other colleagues to determine which constructs should be tested. This is also part of what Fulcher and Davidson (2007, p.86) describe as ‘Engaging with the disciplines, in order to contextualise instruction and assessment in EAP.’ The results showed that 87.3% of teachers supported the actions mentioned, thus suggesting confirmation of their agreement for these methods of enhancing the validity of extrapolations from EAP test scores. Only 1.3% of respondents gave an answer of ‘No’, leaving 11.4% of respondents unsure.

Table 71 Composite results for AQ20 Factor 2 Items AE

Category of response	Summed Frequency of Response	Percent	Cumulative Percent
Yes	276	87.3	87.3
Unsure	36	11.4	98.7
No	4	1.3	100.0
Total	316	100	

Chart 23 Pie chart showing distribution of response to AQ20 Factor 2 (AE)



4.4.13 Composite Likert Scale results for items linked to practices involving analysis and interpretation of the results and scores of EAP tests and assessments

4.4.13.1 AQ22 Factor 1 (DE) - Respondents' views on Criterion/Outcomes-based evaluation and feedback in EAP

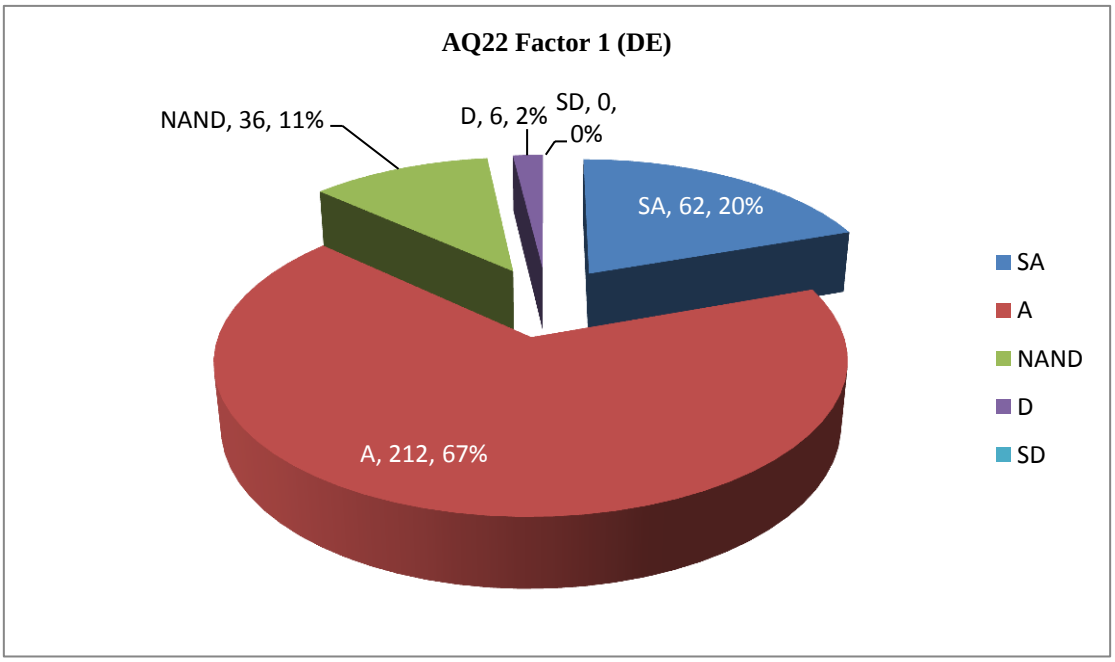
The items included from AQ22 are grouped in terms of their relation to criterion or outcomes-based referencing (Hughes, 2003, p.94). The first question focuses on whether respondents feel that student assessment should be linked to specific learning outcomes and the second question is connected to the attitudes regarding the provision of detailed feedback to other stakeholders, so that scores can be made more meaningful, in the spirit of Assessment for Learning (Biggs & Tang, 2011, p.141).

The findings indicate that 86.7% of respondents either agree or strongly agree with the undertaking of the activities referred to, whilst only 11.4% neither agree nor disagree and just 1.9% disagree.

Table 72 Composite results for AQ22 Factor 1 Items DE

Category of response	Summed Frequency of Response	Percent	Cumulative Percent
SA	62	19.6	19.6
A	212	67.1	86.7
NAND	36	11.4	98.1
D	6	1.9	100.0
SD	0	0.0	100.0
Total	316	100	

Chart 24 Pie chart showing distribution of response to AQ22 Factor 1 (DE)



4.4.13.2 AQ23 Factor 1 (ABDFHJ) and AQ14 (G) - Respondent familiarity with descriptive statistics in EAP assessment

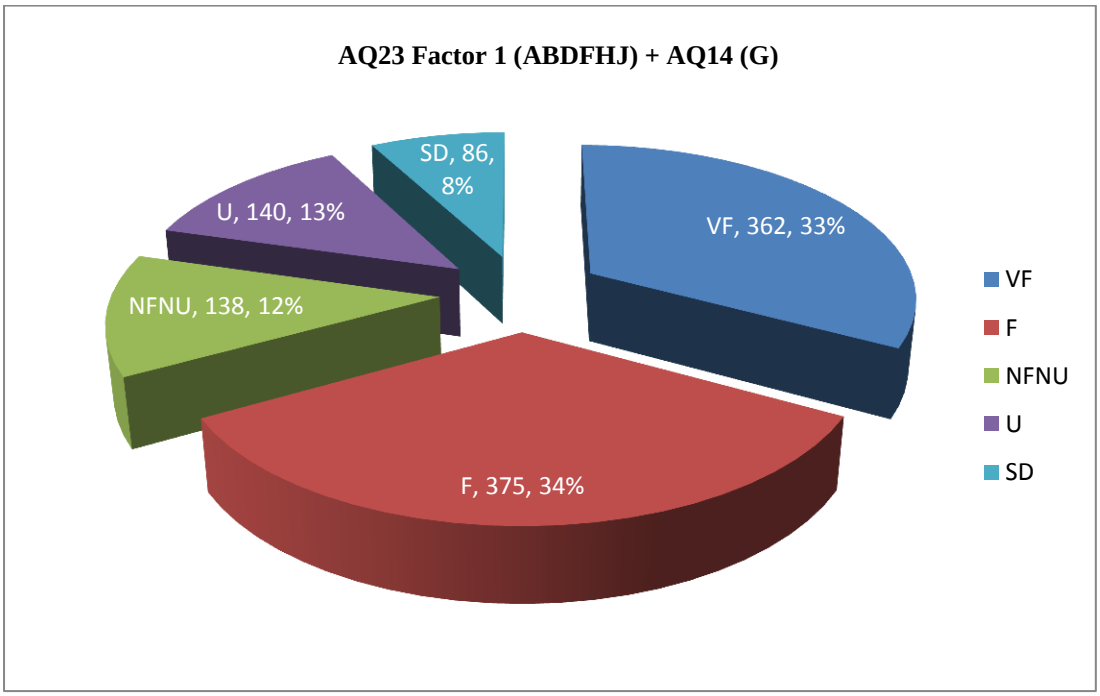
Items in this section of the results were grouped together given that they collectively ask respondents to confirm their level of familiarity with the use of descriptive statistics, in association

with interpreting the results of EAP tests and assessments. Questions relating to specific procedures were asked along with a more general question about familiarity with descriptive statistics. The results demonstrate that 67% of respondents are familiar to some extent with the use of descriptive statistics, with 32.9% being very familiar, to the extent that they have used descriptive statistics themselves. 20.5% were either unfamiliar with the use of descriptive statistics or very unfamiliar. The remaining 12 % of respondents declared that they were neither familiar nor unfamiliar.

Table 73 Composite results for AQ23 Factor 1 ABDFHJ and AQ14G

Category of response	Summed Frequency of Response	Percent	Cumulative Percent
VF	362	32.9	32.9
F	375	34.1	66.9
NFNU	138	12.5	79.5
U	140	12.7	92.2
VU	86	7.8	100.0
Total	1101	100	

Chart 25 Pie chart showing distribution of response to AQ23 Factor 1 (ABDFHJ) + AQ14 (G)



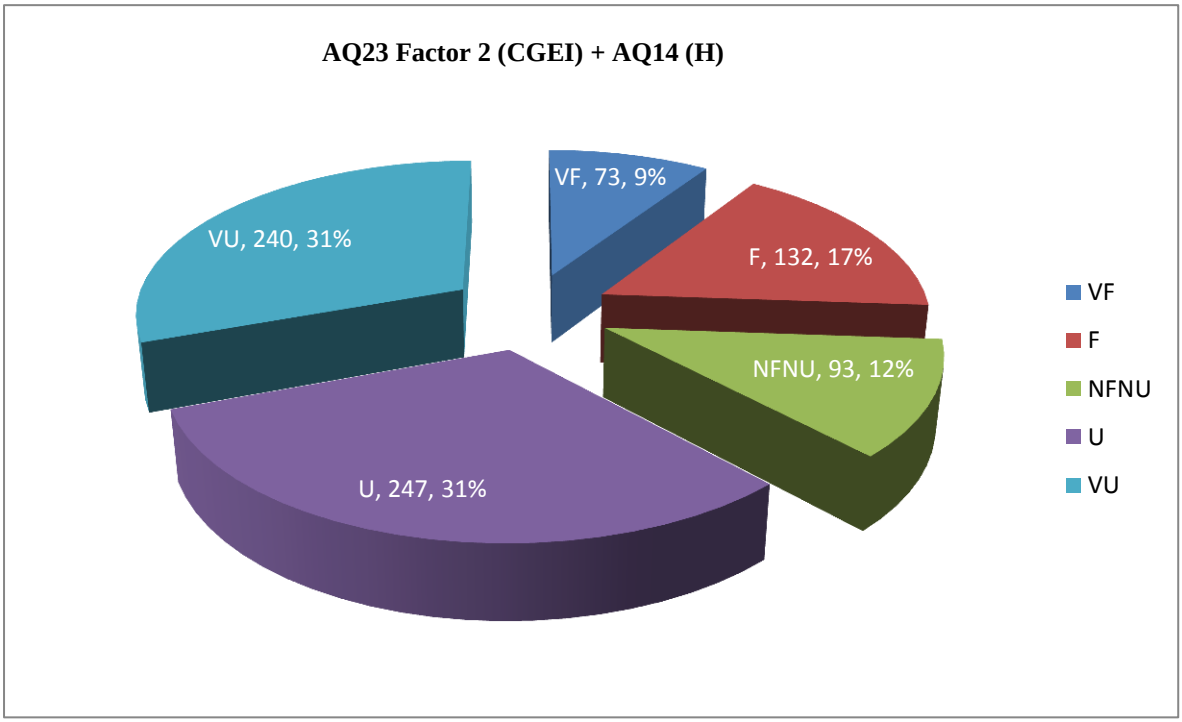
4.4.13.3 AQ23 Factor 2 (CGEI) + AQ14 (H) - Respondent familiarity with inferential statistics in EAP assessment

Respondent familiarity with the use of inferential statistics was the focus in this cluster of items. Questions asked respondents about individual procedures along with a general question relating to overall familiarity. Notably, 62.1% of respondents claimed that they were either unfamiliar or very unfamiliar with inferential statistical procedures, a means of analysis of grades (Brown, 1988, p.115). Only 16.1 % of respondents were familiar or very familiar with inferential statistics and only 9.3% of respondents claimed to have used inferential statistics in their own assessment practice.

Table 74 Composite results for AQ23 Factor 2 (CGEI) + AQ14 (H)

Category of response	Summed Frequency of Response	Percent	Cumulative Percent
VF	73	9.3	9.3
F	132	16.8	26.1
NFNU	93	11.8	38.0
U	247	31.5	69.4
VU	240	30.6	100.0
Total	785	100	

Chart 26 Pie chart showing distribution of response to AQ23 Factor 2 (CGEI) + AQ14 (H)



4.4.14 Composite Likert Scale results for items linked to consideration of ethics in EAP testing and assessment

4.4.14.1 AQ25 Factor 1- Respondents' views on high-stakes EAP assessment dilemmas

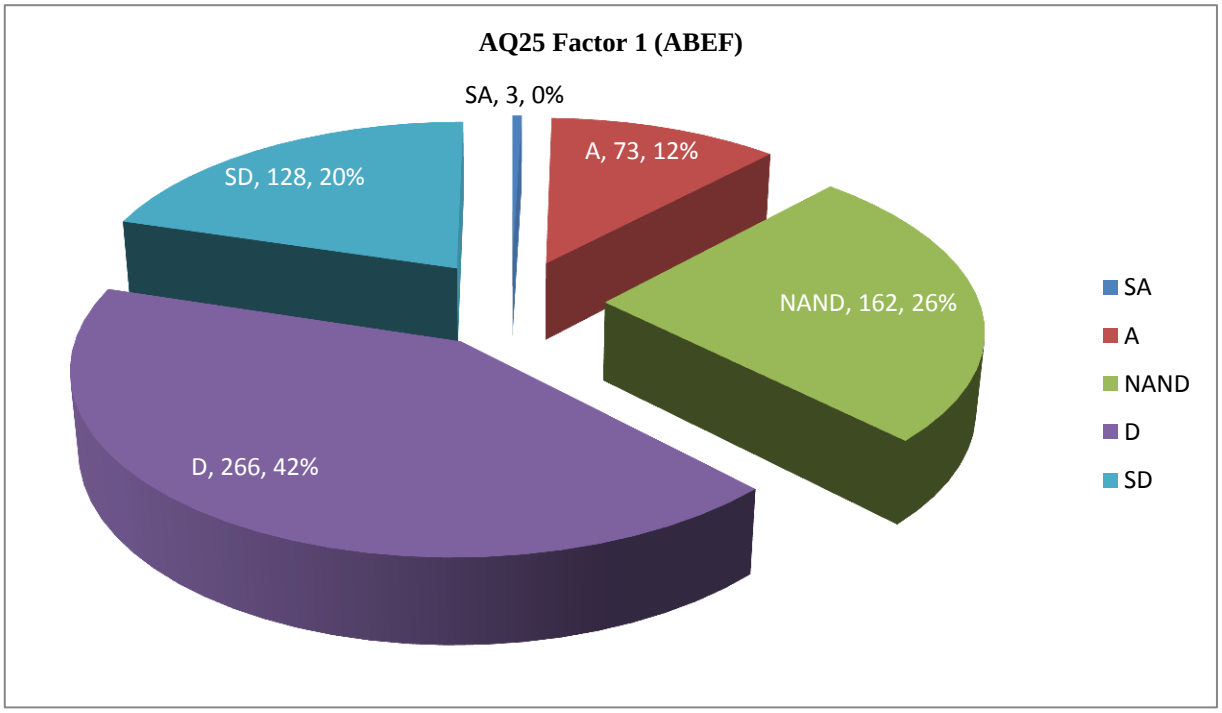
Questions which have been combined in this area of the findings relate to respondents' views on ethical dilemmas. Respondents were asked to comment on a range of situations with complex considerations, given the interaction of different key stakeholders. As items A,B,E and F have all been reversed, due to negative phrasing, agreement indicates respondent support of test-takers' needs and best interests. It could therefore be argued, that agreement also infers an ethical approach, in line with the ethos of Assessment Literacy, which supports the needs of individuals seeking to take tests rather than institutions or other stakeholders (Shohamy, 2001).

The results demonstrate that 62.4% of respondents agree or strongly disagree in response to the dilemmas inferring a stance which supports the needs and best interests of students or test takers. Only 12.1% of respondents recorded their disagreement and arguably adopt a stance which can be argued not to be in the best interest of test takers. 25.4% of respondents neither agreed nor disagreed.

Table 75 Composite results for AQ25 Factor 1 (ABEF)

Category of response	Summed Frequency of Response	Percent	Cumulative Percent
SA	3	20.3	20.3
A	73	42.1	62.3
NAND	162	25.6	88
D	266	11.6	99.5
SD	128	0.5	100
Total	632	100	

Chart 27 Pie chart showing distribution of response to AQ25 Factor 1 (ABEF)



4.4.14.2 AQ26 Factor 1- Ethics in practice in EAP assessment

Items in this section relate to respondents' views on ethics in practice in their own work and employment. Respondents were questioned regarding the extent to which they feel able to avoid unethical practice, the extent to which employers are aware of ethical dilemmas and their views on regulations involving language testing, imposed at governmental level.

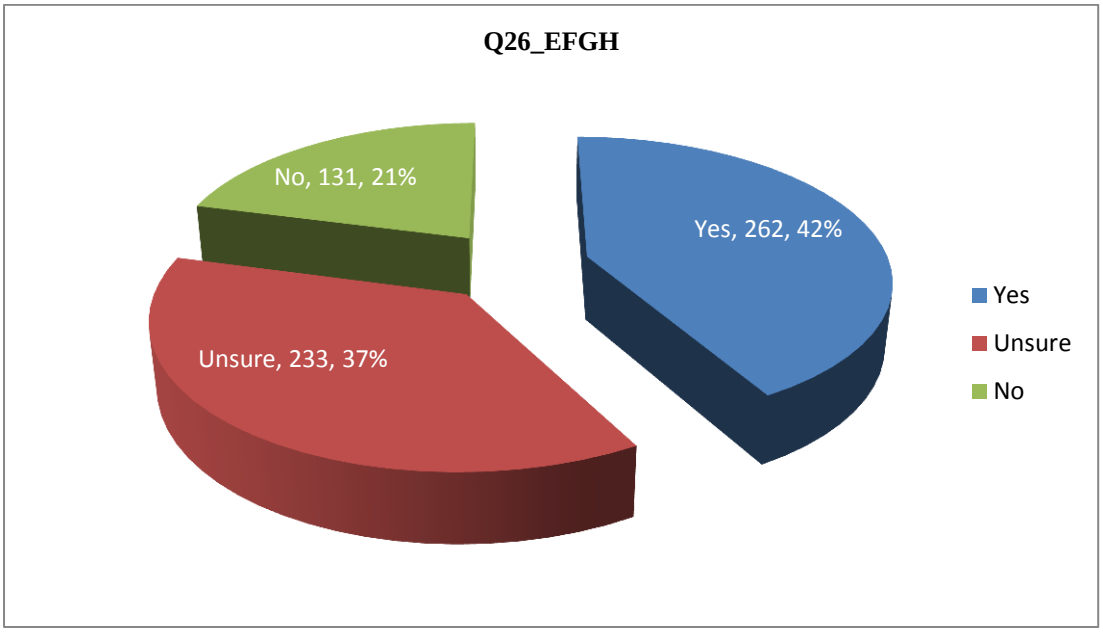
In this series, item G has been reversed so that coding matches that of the other items.

41.9% of respondents answered 'Yes' to this set of items inferring that they feel that their work and employment in association with EAP has not been connected to unethical practices. 20.9% of respondents returned an answer of 'No' which suggests that they do feel that their work and employment has been connected with unethical practice, whilst 37.2% of respondents selected the 'unsure' option.

Table 76 Composite results for AQ26 Factor 1 (EFGH)

Category of response	Summed Frequency of Response	Percent	Cumulative Percent
Yes	262	41.9	41.9
Unsure	233	37.2	79.1
No	131	20.9	100.0
Total	626	100	

Chart 28 Pie chart showing distribution of response to AQ26 Factor 1 (EFGH)



4.4.15 Section conclusion

The aim of this section of the Findings chapter has been to display the results of the questionnaire Likert scales in a manner which has afforded focus on responses to groupings of items which represent particular constructs relating to Assessment Literacy. As these findings relate to common responses to groupings of questions, they can arguably be considered more powerful than responses to individual items. It is believed by the researcher that the subsequent analysis of these findings in Chapter 5 will be amongst the most powerful data for the study, as they represent patterns of response to more than one related-item. Furthermore, the validity of constructs which these interrelated items represent has been tested and enhanced through Factor Analysis. The internal reliability of respondents' views to sets of questions has also been measured through Cronbach's Alpha. As a result, the set of findings which have ultimately been extracted have been filtered to provide more meaningful results for the purpose of analysis in Chapter 5.

4.5 Set A4: Content analysis of open questionnaire responses (See AQ15, AQ18, AQ21, AQ24 and AQ27)

4.5.1 Section introduction

This section presents the results of the open questions drawn from the questionnaire. These open questions were designed to give respondents the opportunity to provide answers without the constraints of predetermined response options. It was hoped that this would offer further scope for respondents to describe their views and practices relating to EAP assessment in a manner which would align with the research questions and their aims to identify evidence of Assessment Literacy, as demonstrated through linkage to language testing research and assessment good practice.

The full results of each open question has been analysed using content analysis procedures, as described in the research methodology in line with models suggested by Gillham (2000c,p.63-66) and Denscombe (2007, p.237). The full set of responses to each open question is available for download through the Appendix in Section 8.4. Content analysis of the raw data has enabled the identification of certain patterns of response which cluster into discernable factors or constructs. These factors or constructs are referred to below in the presentation of findings connected to each open question.

4.5.2 Results from AQ15- Open question relating to training, skills, strengths and weaknesses in EAP testing and assessment.

In total 35 respondents replied to open item AQ15 which invited any additional comments relating to Assessment Literacy in terms of training and skills in EAP assessment. After review of the responses, it was noted that four main factors were observed. Factor 1, contained 11 responses and related to the lack of specific or adequate specialist training in EAP assessment, a concern

highlighted by Sharpling (2002) and Brown and Bailey (2008). Related responses are listed below in Table 77:

Table 77 Absence of specialist or specific training - AQ15 Factor 1

1	Little to no specific training. General ELT experience + DELTA was my entry requirement, and I wish I could have more training.
2	I have not had formal training in assessment and have sometimes felt this lack.
3	I have not attended any specialist training in EAP testing but have picked up some, including through reading. The headings in the questionnaire do not provide me with the best option. I consider I am above average in most of them, but not highly skilled, so I have put average.
4	It's hard to decide if I am skilled in something that I have learnt about 'on the job'. My experiences have shown me that it is incredibly difficult to write good tests and assessments. One has to be very certain of the objectives of the course being taught and then how to assess these objectives accurately and fairly.
5	For a number of years I was responsible for managing a high stakes test but I had no explicit training at the time, and most of the above skills were learned 'on the job.'
6	It's hard to decide if I am skilled in something that I have learnt about 'on the job'. My experiences have shown me that it is incredibly difficult to write good tests and assessments. One has to be very certain of the objectives of the course being taught and then how to assess these objectives accurately and fairly.
7	My MA Applied Linguistics did not focus on this area, and overall my skills have been developed through experience rather than formal training. Assessments I am involved in are related to institution-specific programmes. I therefore find it hard to evaluate my competence, hence the predominance of answers in the 'Average Skilled' column. I am aware of a lack of formal training, however it might be worth mentioning that the course design I am involved in, including assessments, have received good feedback from external examiners.
8	there is no provision in my job for any training in this area and it is much needed
9	My skills have largely come from years of practice and from trial and error.
10	Very little specific formal training in EAP testing and assessment. Sometimes get the impression that it is not in the interests of my institution for staff to know too much about testing for operational reasons.
11	What training? There really isn't any training out there that is specific to EAP testing and assessment, especially if you think beyond large-scale external tests. For in-house assessments, we need to extrapolate from general testing literature which tends to focus too heavily on large-scale tests and not other types of assessment.

The Focus of Factor 2 from AQ15 was the topic of more specific training, where some respondents noted that they had had a particular training experience related to EAP assessment, one even referring to training provided by Weir (see, 1990 and 2005). Five responses load in this category and are listed in Table 78 below:

Table 78 Specific training for EAP assessment -AQ15 Factor 2

1	I have been on the BALEAP working party on language testing for university entry for the past two years and have had to read widely and thought about what can be tested and the validity and reliability of different types of tests.
2	I am also a trained and experienced teacher of maths, which helps me with the statistical aspects of analysis, but am not very familiar with the interpretation of inferential statistics.
3	Began career at Schools Council Language Materials development Project Attended courses at NFER on Educational Measurement Am coordinator and developer of SWELT (Swansea University English Language Test) and STEAP (Swansea Test of English for Academic Purposes)
4	I was part of as testing/assessment group abroad before coming to the UK to teach assessment on an MEd TESOL. I now write tests and check the validity and reliability of other tests.
5	I was trained at Bilkent University in testing and assessment by Cyril Weir and Jake Allsop among others. I worked as Head of Curriculum and Testing Unit there for 2 years (12 years ago).

Factor 3, observed in the data from AQ15, relates to power dynamics, such as those highlighted by Shohamy (2001) and McNamara and Roever (2006), in EAP assessment associated with institutional politics and the balance of control or authority. In total four responses in this grouping were noted as presented in Table 79 below:

Table 79 Power dynamics in EAP -AQ15 Factor 3

1	Sometimes get the impression that it is not in the interests of my institution for staff to know too much about testing for operational reasons.
2	It is hard to rate my skills level as in the previous section when the power dynamics of our set-up limit teacher involvement in assessment design. It is easy to specify areas about which I know nothing at all, but just because I'm not allowed to design assessments doesn't mean I couldn't in a more democratic set-up. Very vexed issues!
3	I am aware of various flaws in the EAP assessments we use, but as a teacher have very little impact on altering them.
4	I am currently working in an institution which has a group of specialist testers. In the five years that I have worked in this institution, the specialist role of those testers has been systematically downgraded in a rather futile attempt to replace diagnostic and summative testing with continuous assessment. This has been an initiative by the present management team, who really do not understand the need for specialist testing and the need for valid sampling.

Finally, in response to item AQ15, Factor 4 groups items which refer to training or the absence of it in connection with MA or Diploma level studies (See Krzanowski, 2001; Roberts, 2001; Sharpling, 2002). There are four responses which group in this area. These have been included in Table 80 below:

Table 80 MA/Diploma-related training for EAP assessment -A15 Factor 4

1	Mainly my training was related to my MA TESOL courses then I attended a few workshops, short courses and forums on testing and assessment.
2	Assessment came up on my Diploma course (Manchester DipTEO) but not on my Masters course in ELT.
3	My MA Applied Linguistics did not focus on this area,
4	On my PG/Diploma in TESOL I studied a language testing module.

4.5.3 Results from AQ18- Open question relating to approaches to EAP testing and assessment design

AQ18 asked respondents to add any further comments which they wished to add in conjunction with EAP test and assessment design. Overall there were a total of 34 responses to this item, in which two main factors were observed.

Factor 1 focuses on the impact of context on the process of EAP test design. The 11 responses, provided in Table 81 below, tend to indicate that the nature of the test design process is reliant upon the particular purpose and setting for which the test is being designed.

Table 81 Impact of context on the EAP test design process- A 18 Factor 1

1	This activity took me quite a long time to think about, and is not an exact science! For example, the need for the test/assessment may come earlier or later on - it might, for instance, emerge naturally from the work the students are doing, so that defining the test purpose may come later on in the sequence. Having said this, there are some things I would probably regard as set in stone - for instance the specifications pre-dating the design of the items, and piloting coming after the items have been created.
2	There are different contexts which could in practice mean different parameters/constraints. there is no consultative stage in which to be advised by academic content teachers
3	The option 'consult past papers' would be only relevant in situations of creating an already-specified test
4	The above sequencing task treats EAP test design as a neatly defined process, which I'm not sure is the case. The research stage is important if the test is new or is introducing new elements which the item writer/test designer is not familiar with. It is less important, surely, if the writers are familiar with the items the test uses and the skills/language areas being tested.
5	Stages 1-8 would happen in an ideal world - in practice, there is often insufficient time and/or resources so test design process is messier and more ad hoc.
6	Some of the options would depend on the purpose - I only wanted to put 5 steps put the survey would not allow me to.
7	I do think these are essential, but also ideal- context and time pressures can impact on how many of these stages take place.
8	Depending on the context some of the above may change order e.g. if there is already a robust test for students going on to a particular subject area, then one would have to add that or combine with stage 1.
9	The order depends on the kind of test I'm developing the context in which the test will be given - difficult to generalise an overall approach which can be ordered into 8 simple stages.
10	There isn't quite a recipe or formula. Each situation is unique with untold numbers of variables.
11	The answer to most of them is actually 'it depends'

The six comments in Factor 2 from AQ18, group around the topic of specificity in EAP assessment with reference to the need to reflect the domain of study pertaining to students studying EAP. Respondent comments are listed below in Table 82.

Table 82 Specificity and domain in EAP assessment - A18 Factor 2

1	I appreciate that it is difficult to design a full range of pre-sessional course assessments, but I believe that students more readily display higher order skills in a subject area they are familiar with and these higher order skills do not necessarily transfer straightforwardly from one context to another.
2	I run a pre-sessional in a generalist university where the range of degree programmes the students are moving onto mean that it is not possible to design tests that relate specifically to students subject areas.
3	For an EAP course intended for students going on to any academic course, the aim should surely be to identify topics of general accessibility. Therefore, I have misgivings about include stage 5 at all and, in fact, might feel that only seven stages are appropriate.
4	I believe EAP assessment has a greater validity if it can replicate what the students will be required to do in future.
5	Tests should be designed with the target language and skills in mind. I have administered tests whose level exceeded the EAP requirements of the students. It is important too to choose academic research examples which are generalised and can be understood by students from a range of disciplines.
6	EAP can cover a wide range of subject areas. I have just left a programme in English for Education, Arts and Law which involved students whose subject majors included English Literature; Early Childhood Education; Primary and Secondary Teaching; Information Technology (Librarianship); Mass Communications and Law. It is both impossible and unfair to even attempt to base a test for those students on one area of subject knowledge.

4.5.4 Results from AQ21- Open question relating to views on/ experience of validity and reliability in EAP testing and assessment

28 responses were returned to open question AQ21, which prompted respondents to add any additional comments relating to Validity and Reliability. The results from this question have been grouped according to three main factors.

Factor 1, which contains 6 responses, relates to respondents' concern to express that validity is highly dependent on factors relating to the particular setting and assessment situation. Table 83 below provides the comments in this grouping:

Table 83 Validity in EAP testing and assessment is highly dependent on other factors - AQ21 Factor 1

1	Again, the answers depend on too many other factors. For example, in theory working as a team is likely to result in a better test, but if the team members are inexperienced, or a member is ignorant but domineering, then the result will not be more valid.
2	In section 19, 'my test' is difficult to define - without seeing your test it is impossible to answer confidently.
3	Much depends on the experiences and personal preferences of the colleagues.
4	it depends on the class
5	Question H will depend on whether it is a test of achievement or proficiency.
6	Validity depends on test purpose. The use of very general terms like EAP testing does not fully take that into account.

The role of domain in the validity of EAP testing is the focus of Factor 2 from AQ21. The five comments in this series, presented in Table 84 below, refer to the need to reflect or sample academic subject domains adequately.

Table 84 Validity & domain in EAP testing and assessment - A21 Factor 2

1	In my experience, validity arises most naturally from conducting a good needs analysis and analysis of the target domain.
2	For me validity comes first- it is about the initial concept and design and whether it is tests something close to what students need to be able to do in an authentic situation and whether it provides evidence that they can do that.
3	Continuous and varied assessment improves reliability but I have often seen, for example, listening tests which really test the students ability to understand the question or texts used in reading tests which are either too genre specific or too far away from the students general area of interest or intended study area.
4	I don't think a pre-sessional assessment can be fully valid unless it relates in some way to the subject the students will be studying at university. I have difficulty imagining a one-size-fits-all EAP assessment.
5	I would aim to sample widely and unpredictably from the syllabus

Comments in Factor 3 from AQ21 refer more specifically to reliability in EAP testing, however it is notable that only three items group in this area (see Table 85 below) and that the reference to reliability is quite diverse.

Table 85 Reliability in EAP testing and assessment - AQ21 Factor 3

1	I think reliability could be an issue if the test is very specialised, as some markers may not feel 'qualified' to comment on students' work, and there may be variability in what is being looked for. However, there is nothing here that good rater training could not solve. Negotiation and arriving at a shared understanding of what is being looked for seem to be the key areas for achieving reliability.
2	Reliability is more a quality of the finished product- it is about the technical aspects of question writing and exam setting and administration.
3	team work with other tutors and meetings are vital to avoid subjectivity and be sure of benchmarks

4.5.5 Results from AQ24- Open question relating to practices involving analysis and interpretation of the results and scores of EAP tests and assessments.

In response to AQ24, which allowed respondents to add additional comments connected to the analysis and interpretation of EAP tests and assessments, 18 responses were provided and two principal factors were returned, the first factor related to the impact of IELTS or other major standardised tests at local level. The four comments connected to AQ24 Factor 1 are presented below in Table 86:

Table 86 The impact of IELTS and other standardised tests at local level AQ24- Factor1

1	We have the situation where the in-house home-made (mini-IELTS copycat) test is deemed to be considered by the uni as equivalent to IELTS and management now seem keen to undermine the reliability of IELTS since candidates were allowed to repeat the test ad lib. They can't quibble with the validity of IELTS as they have tried to copy it, but they are comparing students' results in the home-made test and IELTS to suggest I'm not quite sure what - sometimes that IELTS gives different scores to same student on different occasions, sometimes that the IELTS score is much the same as the score on the home-made test. Hope that makes sense!
2	Although it may be seen to be 'impractical' to convert EAP test results into IELTS equivalent it is sometimes very necessary to enable colleagues working within an admissions role to understand a student's level of language - IELTS is far more widely understood across the UK HEI sector than any other test.
3	22f Other people throughout the university should ask us more about, e.g., IELTS, but they don't! 22h This could reveal that the test is at fault, or that the students need more work in that area.
4	Despite all those intensive language tests and English language requirements (TOEFL, IETLS, PCU, and locally MUET), the reality is that many 'slip through the net', giving rise to a situation of extreme mixed-ability classes - and no money or time for proper streaming. The good ones will then start skipping class; the weak ones end up doing exercises on the simple past versus the present perfect! Somewhere in all of this, I am sure, there is room for EAP testing/assessment of the type envisaged by EAP departments but right now, we simply work with what we have.

The three responses which can logically be grouped for AQ24 Factor 2 are connected to the use of statistical procedures in EAP test score analysis. These comments are included in Table 87 below:

Table 87 Use of statistical procedures in EAP test analysis- A24 Factor 2

1	Item facility rating should be included
2	I used IRT rather than classical item analysis - but it was a while ago and we had an expert on hand to help!
3	Year-on-year basic statistical analysis is a really good idea however to measure how two groups compare.

4.5.6 Results from AQ27- Open question relating to consideration of ethics in EAP testing and assessment.

The final open item in the questionnaire asks respondents to add any additional comments connected to ethics in EAP testing and assessment. 22 responses were provided to this question. The first factor which has been observed in responses is connected to respondents' statements of ethical position. The eight comments in this factor have been listed as they each represent a view or ethical stance in association with an institution or individuals who are involved with or affected by EAP testing. Such concerns are linked in particular to EAP by Benesch (2001, p.60).

Table 88 Statements of ethical position - A27 Factor 1

1	This depends very much on University Policy… But some places are just so rigid in their enforcement of rules.
2	Teachers who write /correct EAP exams should be careful; if they are not students will eventually find out and start to mistrust the system an d the teachers
3	I feel that courses in EAP, ESP, etc. should play a supportive role and that small-group tuition/guidance based on real assignments, etc. will over time yield better results. Of course, the cost implications are such that this is unlikely to happen ever. It's easier to adopt a factory-like system with batches we process and 'rejects' that are left behind.
4	bullying and unskilled assessors should be held to account given the gravity of final assessment particularly regarding borderline cases
5	I have been involved recently in a project which assessed the performance of a group of international staff in a lecture context, and this raised huge ethical issues.
6	Ethical considerations are of paramount importance in language testing. My own view is that we should establish a humane testing culture which is constructive and collaborative, rather than repressive. There is a school of thought (Shohamy, MacNamara, etc.) which is looking more closely at the social impact of testing and the power of tests, and this strongly informs my practice. I think my own personal experiences and identity help me to see discrimination a bit more easily than is the case for some others, so I see my role as educative in this area.
7	I have written tests for our own EAP foundation course, also for our in house IELTS equivalent for Chinese students who take our test in China, however, until very recently, none of the wider issues or in fact any knowledge about language testing was known about by me. I recently did my diploma so have a little more understanding now.
8	Pre-sessional course managers are essentially magicians whose job is to deliver a predetermined number of students to departments while using a smokescreen of procedures to give the satisfying illusion that students have been rigorously screened for their ability to cope linguistically with follow-on courses. The university cares about ethics only if there is a real danger of being used.

Factor 2, relating to AQ27, relates to respondents who felt unable to answer questions relating to ethical dilemmas due to the dependence of additional factors and the need for additional information. Table 89 below shows four comments where respondents have noted their difficulty in answering certain questions for the reasons described above.

Table 89 responses to ethically-related questions are dependent on other factors- A27 Factor 2

1	My 'Unsures' are largely because the answer is 'it depends'.
2	The answers to many of these questions depend on individual circumstances - e.g. for 27 i if the student can prove identity in another way he/she should be allowed to take the test.
3	25c - depends when in the test this happened
4	These questions are too complex to answer in this way.

4.5.7 Section conclusion

The results from this section of the Findings chapter have allowed the identification of clusters of response to open questions, where respondents were given the opportunity to add additional views, which could then be considered in terms of what they reveal about Assessment Literacy. The function of these questions served to provide an opportunity to respondents if they did not feel able to express themselves fully through the closed questions within the questionnaire. It should be noted that the relative impact of these findings needs to be considered, given that response to these open questions was not compulsory. Consequently, the number of responses was reduced. Nevertheless, the patterns of response identified will still be useful to consider, with the above context in mind, during the process of analysis in Chapter 5.

4.6 Set B1: Interview data- sample population description

4.6.1 Section introduction

In this section of the Findings, results will be presented relating to the features of the overall interviewee sample population. In order to be able to refer to a population with similar characteristics to the population which was identified at the questionnaire stage of research, respondents who supplied their email address through the questionnaire were contacted for participation in the interview stage. Approximately 35 questionnaire respondents were contacted and finally 25 interviews were conducted.

4.6.2 Interview population characteristics and stratification in line with questionnaire respondents

In the process of identifiying appropriate interviewees from the population of questionnaire respondents in Population B, efforts were made to practice a form of stratified sampling (Dörnyei, 2007, p.97). As a result, potential interviewees were identified in manner which would allow a

similar spread of features such as gender, age, years of experience, country of practice and institution type. Table 90 below describes the interviewee population characteristics.

Table 90 Characteristics of interviewees IA-IY

Interviewee Coding	Gender	Age Group	Years of experience	Institution Type	Nationality	Country of Practice
I_A	F	35-44	5 years +	University	UK	UK
I_B	F	45-54	10 years+	University	US	UK
I_C	M	45-54	10 years+	University	UK	UK
I_D	M	22-34	2-5 years ago	University	French	UK
I_E	F	55-64	20 years+	University	UK	UK
I_F	F	35-44	2-5 years ago	Partnership	UK	UK
I_G	F	55-64	10 years+	University	UK	UK
I_H	F	22-34	2-5years	University	US	US
I_I	F	65+	20 years+	University	UK	UK
I_J	M	35-44	5 years+	University	UK	China
I_K	M	22-34	2-5 years	University	UK	UK
I_L	F	54-64	20 years+	Partnership	UK	UK
I_M	F	22-34	0-23 months	University	UK	UK
I_N	F	35-44	10 years+	Partnership	UK	China
I_O	F	45-54	10 years+	University	Australian	Australia
I_P	F	35-44	5 years +	Partnership	UK	UK
I_Q	F	55-64	20 years+	University	UK	UK
I_R	F	35-44	10 years+	University	UK	UK
I_S	F	45-54	10 years+	University	UK	UK
I_T	M	35-44	10 years+	University	UK	UK
I_U	M	45-54	5 years+	University	UK	UK
I_V	F	55-64	0-23 Months	University	Irish	UK
I_W	F	45-54	10 years+	University	UK	UK
I_X	M	55-64	5 years+	University	UK	France
I_Y	M	45-54	10 years+	University	UK	UK

Whilst the features of the final population were partially dependent on the individuals who agreed to participate, it is notable that the population does share key similarities with that of the questionnaire stage.

With regard to gender split, 32% of the total population of interviewees were male in comparison to 34.8% of the total number of questionnaire respondents in Population B, which was the final

relevant questionnaire population identified. In terms of age groupings, Table 91 below shows the similarity in spread of age groupings across both the interview and questionnaire stages of data collection.

Table 91 Comparison of interview and questionnaire populations by age grouping

Age grouping	Number of questionnaire respondents (Population B)	Percentage	Number of interviewees	Percentage
22-34	4	16	20	12.7
35-44	7	28	47	29.7
45-54	7	28	50	31.6
55-64	6	24	37	23.4
65+	1	4	4	2.5

When years of service in EAP is considered, Table 92 below also shows that a similar balance of age groupings was maintained across both questionnaire and interview respondent groups.

Table 92 Comparison of interview and questionnaire populations by years of service

Period of service	Number of questionnaire respondents (Population B)	Percentage	Number of interviewees	Percentage
0-23	11	6.96	2	8
2-5	31	19.62	4	16
5+	32	20.25	5	20
10+	56	35.44	10	40
20+	28	17.72	4	16

The balance of institutional affiliation has also been maintained in the interview data when compared to the questionnaire, with 16% of interviewees stemming from university partnerships in comparison to 14% of interview respondents.

As far as nationality and country of practice is concerned, the interviews showed a nationality split of 80% for UK with 20% of interviewees representing nationalities from the rest of the world. The results of the questionnaire showed that 73% of respondents identified UK as their nationality with 27% stemming from countries outside of the UK.

For country of practice, the interviews showed that 80% of interviewees have most recently worked in EAP in the UK. 20% of interviewees have worked most recently in countries across the rest of the world. For the questionnaires, 34.8% of respondents have most recently worked in EAP overseas, whilst 65.1% have worked most recently in EAP in the UK.

4.6.1 Section conclusion

After consideration of the features of the population and the measures which have been taken to stratify the population of interviewees, it has been judged by the researcher, and through corroboration by a colleague, that the interviewee population represents a suitable population to proceed with further, for the findings of this area of the research.

4.7 Set B2: Interview interactional profiles

4.7.1 Section introduction

In addition to the reporting of population-related statistics, which have already been presented, and the clustered results drawn from a content analysis which will appear in Set B3 of the Findings chapter, it is also useful to consider the nature of the set of interviews in terms of the interactional

encounters which they represent (Heigham & Croker, 2009, p.193) and the individual positions embodied by each of the cases or interviewees (Drever & Scottish Council for Research in Education., 2003, p.72).

4.7.2 Structure and results of interactional profiles

The interactional profiles which were compiled for the set of interviews used in this research were completed using notes taken by the researcher during the process of the interviewing, in conjunction with demographic data, the duration of the interview and observations relating to the relative status of the interviewer and interviewee during the interview. The level of engagement of the interviewee, as discerned by the interviewer, was also noted.

A sample of three profiles relating to the interviews undertaken is provided below. The full set of profiles can be found in the Appendix of this document in section 8.8 on page 468.

Table 93 Interactional profiles IA, IG & IY

Interviewee Code	I_A	Interview Mode	Skype	Institution Type	University
Country of Practice	UK	Gender	F	Age group	35-44
Length of Interview	19 Minutes 23 Seconds			Years of practice	5 years +
Features of Interaction	Interviewee/Interviewer Status	Peer-to-Peer Moderate acknowledgement of interviewer having higher level of experience.			
	Interviewee Engagement	Engaged with subject but notably inexperienced in some areas			
	Other Comments	High level of enthusiasm but does not consider self as an expert. Time-pressured interview			

Interviewee Code	I_G	Interview Mode	Face-to-Face	Institution Type	University
Country of Practice	UK	Gender	F	Age group	55-64
Length of Interview	17 Minutes 17 Seconds			Years of practice	10 years+
Features of Interaction	Interviewee/Interviewer Status	Peer-to-Peer / Junior-to-Senior Tone of interaction indicated interviewee's wish to appear senior			
	Interviewee Engagement	Neither engaged nor disengaged			
	Other Comments	Interviewee very confident of their own approach despite its flaws which were acknowledged by interviewee.			

Interviewee Code	I_Y	Interview Mode	Skype	Institution Type	University
Country of Practice	UK	Gender	M	Age group	45-54
Length of Interview	32 Minutes 25 Seconds			Years of practice	10 years+
Features of Interaction	Interviewee/Interviewer Status	Junior-to-senior -despite approachability of interviewee their level of skill and experience indicated seniority to interviewer			
	Interviewee Engagement	Very engaged			
	Other Comments	The interviewee appeared very highly skilled and experienced in the field			

The accuracy of the interactional profiles was also verified through the use of a moderator who reviewed the researcher's analysis of a representative sample of interactional profiles.

4.7.3 Section conclusion

Through consideration of the interactions which have taken place, in the interviews which were undertaken, it will be possible to develop a closer view of how the conversational encounters have been constructed and whether the individual interviews raise any concerns, given that the manner in which the interviewee and interviewer interact may affect the type and tone of response given by the interviewee and the manner in which this is interpreted (Wooffitt & Widdicombe, 2006). Further examination of the interactional profiles and what the researcher perceives that they reveal will be undertaken in the Analysis and Discussion section of the thesis.

4.8 Set B3: Content analysis of interview question responses BQ1-BQ18

4.8.1 Section introduction

Results of the interview stage of the research have been provided below in a series of tables showing the main clusters of factors or components which emerge from each respective interview question. The term, 'factor' will be used to refer to patterns of meaning. Only those factors have been reported where three or more interviewees refer to a connected point. Below each table of factors a table showing three examples of each factor has been provided to illustrate the points being made and allowing the voices of the interviewees to emerge (Heigham & Croker, 2009).

Clusters of factors were first identified by the researcher and then corroborated with the assistance of a colleague with experience in the field, in order to substantiate or reject perceived observations.

In order to contextualise the findings described in this section, groupings of meaning which have emerged from the interviews have been organised according to the overarching aspects of Assessment Literacy which were identified as key in Table 31 on page 169. Table 31 links key areas of EAP Assessment Literacy to the items or sections within the interview protocol where they have been included. As mentioned previously, these particular areas of EAP Assessment Literacy were established as significant to the research questions during the Literature Review and have been summarised in Section

In each of the charts which follow, reference to a factor or category of meaning, in one of the interviews is highlighted by marking a corresponding cell with '/' and highlighting the corresponding cell in blue.

4.8.2 Results of content analysis for items linked to training, skills, strengths and weaknesses in EAP testing and assessment

4.8.2.1 Major factors identified from interview Question BQ1

The first series of factors presented in this section, taken from responses to AQ1, relate to an initial open question at the very start of each interview. The results of this question have been organized in this section alongside other items which relate more specifically to training and skills, for orgnaisational convenience.

The aim of this section was to allow respondents the opportunity to add any general comments before more specific questions related to EAP assessment were posed. As shown in

Table 94 below, the most notable pattern and clustering of responses which emerged from Factor 1, related to respondents who did not feel the need to comment. It seemed that this was largely in the expectation that they would have a subsequent opportunity to provide additional comments. Examples of the factors from BQ1 are listed below in Table 95.

Table 94 Factors identified in BQ1 - Interviewee opening free comments

Interviews ↓	Factors→			
	(1)	(2)	(3)	(4)
I_A	/			
I_B				
I_C		/		
I_D				
I_E			/	
I_F				
I_G	/			
I_H	/			
I_I	/			
I_J	/			
I_K	/			
I_L		/	/	/
I_M	/			
I_N				
I_O	/			
I_P		/		
I_Q	/			
I_R	/			
I_S	/			
I_T		/		
I_U	/			
I_V				/
I_W			/	/
I_X				
I_Y	/			

Table 95 Examples of factors identified in BQ1

Factor number	Description of factors identified in interviews	Examples drawn from interviews (interviews coded using letters in brackets e.g. '(A)' denotes 'Interview A')		
1	No comment	(Q) No, not really	(O) No, that's okay	(M) We might as well carry on I think.
2	Lack of validity/reliability	(C) we've just got to be honest and say for example, this test isn't valid anymore	(P) My focus is authenticity in writing tasks and creation of positive washback from that so the bee in my bonnet is how discursive essays are a bit pointless.	(T) Having just come out from more or less a moderation meeting and…where we didn't agree with each other and you know, referring to the discrepancies on it. It's not easy.
3	Language & skills balance – Integrated Skills	(E) Well, no. Well, probably I feel like the assessment sometimes tends to focus too much on the skills and not enough on the language.	(L)You'll probably end up not really assessing the kinds of skills that you want to assess.	(W) I'm very much in favour of achievement testing at the end of EAP courses … so you know reflecting the skills that they've been learning on the course
4	Issues associated with IELTS & TOEFL	(L) To ensure reliable assessment, you'll more likely to get reliable assessments statistically speaking if you use a test which requires objective answers like multiple choice or one-word answers that basically the IELTS format for assessment.	(V) And I think IELTS doesn't necessarily prepare them for academic study. You know like they can get quite a good score and still be a little bit lost. Is it because it's so specialised, you know?	(W) Rather than using proficiency test and you know, a sort of IELTS language-type test.

4.8.2.2 Major factors identified from interview Question BQ2

The results which follow, drawn from BQ2 and displayed in Table 96, cover the subject of interviewee training and skills development. This question aims to examine how interviewees have developed Assessment Literacy skills associated with EAP assessment. This area aligns with the concerns as expressed by supporters of Assessment Literacy such as Popham (2009) and Stiggins (1991, 1995) regarding the need for educationalists to develop a skill set which is specifically related and relevant to assessment.

Table 96 Factors identified in BQ2 - Interviewee training and skills development

Interviews ↓	Factors→				
	(1)	(2)	(3)	(4)	(5)
I_A	/	/			
I_B	/				
I_C	/		/		
I_D	/		/	/	
I_E	/	/			/
I_F	/				/
I_G	/	/			/
I_H	/			/	
I_I	/				
I_J	/				
I_K			/		/
I_L		/			
I_M	/				
I_N	/	/			/
I_O	/		/		
I_P	/				
I_Q		/		/	
I_R	/	/			
I_S	/	/			
I_T	/	/	/		
I_U	/		/		
I_V					
I_W					
I_X	/				
I_Y	/		/		

The clearest pattern and grouping of comments shown in Table 96 above stem from Factor 1, where interviewees commented that their skills in EAP assessment and testing had been developed either 'on the job' or through professional practice. This observation was also made by Sharpling (2002). This Factor 1 can also be seen to overlap with both Factors 4 and 5, which refer to the development of skills either through working with colleagues or through in-classroom experience. Factor 2 refers to respondents who were able to acknowledge forms of specific assessment-related training. Below, Table 97 provides a series of illustrative examples of each of the factors noted.

Table 97 Examples of factors identified in BQ2

Factor number	Description of factors identified in interviews	Examples drawn from interviews (interviews coded using letters in brackets e.g. '(A)' denotes 'Interview A')		
1	Informally or on the job	(B) On the job and some reading but not a lot.	(I)I think probably through being asked to teach on such courses and being given targets for my students.	(T)I've developed them just as I've gone along really as much as by using them, by trying to assess the tested work and so on.
2	Through training courses	(E)In my MA, in Applied Linguistics, I took a testing course. That's what I did at the beginning.	(Q) Through the training, I've done myself. I mean I did a testing option as part of my masters, which made me think about testing, looking at other pre-existing texts	(S)Well, definitely through my own studies, obviously. I've been doing my Master's Degree.
3	Lack of formal training	(C) I have never had any personal training how to do that.	(T) I've never had formal training in testing at all.	(Y) I haven't undergone any specific language testing training as such.
4	Through considering student needs or learning requirements	(D) just by looking at what students need to achieve at the end and then working backwards	(H) You don't know until you're there in the classroom with the students and they all are different. So it really is the context, absolutely.	(Q) the way that I've developed things here is looking at the skills that are required to be tested, and then trying to have a sort of integrated exam format
5	Through working with groups of colleagues	(E)And then teaching on the CALS pre-sessional. And especially when I did that, the head of it was also a testing teacher and was head of testing as well	(K) Through work with course leaders who have shared their thoughts, and then experimenting with things like wanting to try it myself.	(N) I learned most of it by talking to my peers because at that time, I was working in study groups with different people

4.8.2.3 Major factors identified from interview Question BQ3

The details presented in this part of the findings, collected from BQ3 and presented in Table 98, provide patterns of response concerned with interviewee strengths in EAP assessment. The aim of this question was to identify key skill areas associated with Assessment Literacy.

Table 98 Factors identified in BQ3 – interviewee strengths in EAP testing and assessment

Interviews ↓	Factors→		
	(1)	(2)	(3)
I_A			
I_B			
I_C			
I_D			
I_E			
I_F			
I_G			/
I_H			
I_I		/	
I_J	/		
I_K			
I_L			/
I_M		/	
I_N			
I_O			
I_P			
I_Q	/		
I_R		/	
I_S	/		
I_T	/		
I_U			/
I_V	/		
I_W			/
I_X			
I_Y		/	

As Table 98 above shows, a similar incidence of grouping has emerged from each of the three factors identified, in the areas of gauging student learning needs, with an emphasis on a formative approach (Biggs & Tang, 2011, p.64; Knight, 1995), teachers' academic background in terms of subject knowledge or ESAP, and the act of producing test materials, as a part of test design. Examples which illustrate the factors in Table 98 are listed below in Table 99.

Table 99 Examples of factors identified in BQ3

Factor number	Description of factors identified in interviews	Examples drawn from interviews (interviews coded using letters in brackets e.g. '(A)' denotes 'Interview A')		
1	Gauging student needs	(J) One of my strengths is that I find it quite easy to gauge people's level quite early on.	(Q) Well I think I understand our own students' needs quite well	(S) I'm very flexible in terms of, I won't say, "Well, this is how we assess." This kind of problem or this kind of learning. Because actually, other students might need something slightly different.
2	Teachers' academic background or knowledge of ESAP in practice	(I) I think because I've got an academic background myself, I can say to the students, "You need to do this for this reason."	(M) My strengths are when it is an ESAP situation that I am familiar with.	(Y) I have a good knowledge of the field in terms of you know the main specific language. I'm quite aware of the sort of differences in sort of academic writing conventions that I used from one department to another.
3	Producing test materials	(G) Having written quite a lot of tests, you know, I've written for IELTS, for example, that's clarified a lot my mind.	(U) I'm quite good at producing current assessment which is interesting and can lead onto essays that are easy…if I give students something to write about.	(W)I'd like to think I'm good at item writing and constructing tests.

4.8.2.4 Major factors identified from interview Question BQ4

Consideration of the results from this next item in the interview, compiled from BQ4, shows interviewees' opinions on their weaknesses in EAP asssessment. The aim of this question was to reveal areas of Assessment Literacy requiring specific development. The key results of this question are revealed in Table 100 below.

Table 100 Factors identified in BQ4 – interviewee weaknesses in EAP testing and assessment

Interviews ↓	Factors→
	(1)
I_A	/
I_B	
I_C	
I_D	
I_E	
I_F	
I_G	
I_H	
I_I	/
I_J	
I_K	/
I_L	
I_M	
I_N	/
I_O	
I_P	
I_Q	
I_R	/
I_S	
I_T	/
I_U	/
I_V	
I_W	
I_X	
I_Y	

As Table 100 indicates, the only key factor which showed a marked cluster of responses across a number of interviewees related to a perceived lack of competence or experience in EAP testing and assessment. Brown and Abeywickrama, (2010, p.xi) acknowledge the sense of inadequacy which teachers sometimes experience when considering testing and assessment. Examples of responses drawn from the interviews which highlight perceptions of limitations in experience or competence are presented in Table 101 below.

Table 101 Examples of factors identified in BQ4

Factor number	Description of factors identified in interviews	Examples drawn from interviews (interviews coded using letters in brackets e.g. '(A)' denotes 'Interview A')		
1	Lack of competence or experience	(I) I don't regard myself as a very competent or consistent marker.	(K) I don't think I have fully worked out a coherent conception of what I'm aiming to do with testing, how it all fits into an integrated whole.	(U) Weaknesses, well the, my biggest weakness is I'm just not very good at it.

4.8.2.5 Major factors identified from interview Question BQ5

Information drawn from this following group of items, summarised from BQ5 and displayed in Table 102 below, is concerned with the biggest challenge in EAP assessment which interviewees claim to have encountered. Again it was envisaged that interviewees'responses to this item would reveal areas of potential skill development connected to Assessment Literacy in EAP assessment.

Table 102 Factors identified in BQ5 – interviewee most challenging aspect of EAP testing and assessment

Interviews ↓	Factors→		
	(1)	(2)	(3)
I_A	/	/	
I_B			
I_C			
I_D	/		
I_E		/	
I_F			
I_G			
I_H			
I_I			
I_J		/	
I_K			
I_L			
I_M		/	
I_N		/	
I_O			
I_P		/	
I_Q			/
I_R			
I_S			/
I_T	/		
I_U	/		
I_V	/		
I_W			
I_X			/
I_Y			

The factors which arose from question BQ5, as seen in Table 102, which showed the most prominence were associated with interviewee concern to undertake assessment well, or 'to get it right' and to assess EAP in a manner which reflects students' real needs in the university context. Other factors can be linked to communicative language testing through assessing what students require at university (Weir, 1990) and dealing with the challenges presented by specific purposes, such as those described by Swales, (1985). Examples highlighting the factors identified are provided in Table 103 below.

Table 103 Examples of factors identified in BQ5

Factor number	Description of factors identified in interviews	Examples drawn from interviews (interviews coded using letters in brackets e.g. '(A)' denotes 'Interview A')		
1	Knowing what to test/ getting it right	(A) Ambiguity about what I'm supposed to be getting from the student	(D) Knowing what exactly to test because there are so many different things	(U) It's just, it's just getting out an assessment which is at the right level and produces a decent spread.
2	Assessing what students really need to do at University	(A) Mismatch between what the University wants through IELTS and fact that IELTS is not particularly academic.	(J) By far, and away the most challenging thing is to make sure that the testing is going to appropriate for what they need to use it for.	(N) They can do that test but it doesn't mean anything in the real world. So I think with EAP testing, you have to make sure it's got some kind of significance somewhere
3	Dealing with diverse/particular subject areas	(Q) think it's the whole concept of what is EAP, because when you're looking at so many different genres, and you've got possibly people from all different subject backgrounds, how do you actually find a test which would suit all	(S) So I would say that probably the most challenging thing is when you…what I've just been doing really is marking a whole lot of essays which are very Finance orientated … I get quite frustrated at the thought that actually, I don't have that level of knowledge they've got	(X) Sometimes the subject that they're writing about. (Laughter) But when in terms of the language it's, I think it's always a challenge to read what somebody else is trying to express

4.8.3 Results of content analysis for items linked to approaches to EAP testing and assessment design

4.8.3.1 Major factors identified from interview Question BQ6

The data collected from this section of the interview, selected from the results of BQ6 and shown in Table 104 below, focuses on comments provided by interviewees in connection with processes that they have experience of employing in designing EAP tests and assessments. The aim of this section was to identify the extent to which interviewees follow set procedures in EAP test design in a manner which facilitates principled test design, such as that suggested by practitioners like Fulcher and Davidson (2007).

Table 104 Factors identified in BQ6 – interviewee views on EAP test and assessment design process

Interviews ↓	Factors→					
	(1)	(2)	(3)	(4)	(5)	(6)
I_A	/					
I_B	/	/				
I_C	/					
I_D						
I_E	/					
I_F	/		/			
I_G	/		/			
I_H						
I_I	/	/				
I_J				/		
I_K	/				/	
I_L					/	
I_M	/		/			
I_N		/				
I_O						/
I_P						/
I_Q			/			
I_R						/
I_S	/		/			
I_T	/					
I_U				/		
I_V				/		
I_W					/	/
I_X				/		
I_Y	/					

A number of the observed factors above in Table 104, do demonstrate a number of smaller clusters of commonality between interviewees, showing principled and organised approaches to EAP test design. This includes the process of needs analysis or the description of particular procedures. However, the most common grouping of response to this item showed that 12 interviewees reported no identifiable procedure. Despite concerns expressed by some regarding the unnecessary rigidity of set processes (Jafarpur, 2003, p.72), this unstructured approach could risk leaving

important questions unanswered (Fulcher, 2010, p.94) in the practice of EAP test and assessment design. The examples shown below in Table 105 present an overview of interviewee response related to each of the main categories of response identified.

Table 105 Examples of factors identified in BQ6

Factor number	Description of factors identified in interviews	Examples drawn from interviews (interviews coded using letters in brackets e.g. '(A)' denotes 'Interview A')		
1	No identifiable procedure	(C) I'm aware there isn't a process.	(M) I would like to be able to give you a clear set of steps but I feel it is a bit more of an organic process.	(F) (Set process?) Not at the moment. No.
2	Looking at what happened/was done in previous years	(B) Honestly looking at previous years, what we've already done is usually the baseline	(I) I mean I suppose the conventional answer, is you'd look at what had come before	(N) I'd look at previous models… Looking at what hasn't worked in the past
3	Start by looking at students skills and needs	(F) I'd start thinking about which sort of sub-skill area it was obviously and the level of the students in the class.	(S) So I would first of all make sure that I knew what those needs were. That I knew what they needed to accomplish and where they needed to go.	(Q) Yes, I suppose first of all starting with what needs to be tested, and I would usually do that on a skills basis.
4	Interviewee describes a set process or format	(J) Usually two academic pieces of writing one of which would be a report or research-based paper, one of it would be an essay based paper…	(T) in the situation I've done EAP before, there's has been a set process and I've you know, used what was given and supplemented my own…just supplemented my own judgment and intuition.	(U) So I find the listening text first. And then try and find the reading text which are, which are thematically connected
5	Assessment designed as part of or with reference to course design	(K) I guess actually the assessment design is part of the course design.	(L) I'll draw up the specifications at first. Well, I need to know certain factors in which the test was meant to be used	(W) I'm writing the outline and the assessment at the same time. And the task would be very linked to the outcomes.
6	Alignment with outcomes/goals	(O) Constructive alignment and so on. So, working out what I want the outcome to be and then making an assessment that would fit those outcomes	(P) I suppose that the main principle will be to keep in mind their end goal and think about how close or far away people are from that and try to base assessment on appropriate steps towards the end goal	(R) We had a sort of set formula where we were testing different things and it was to do with the outcomes of the courses.

4.8.3.2 Major factors identified from interview Question BQ7

The results which emerge from this section, taken from BQ7 and presented in Table 106, show responses connected to practices employed by interviewees in identifying and selecting suitable test items for their assessments.

Table 106 Factors identified in BQ7 – interviewee views on how to identify questions to include in EAP tests

Interviews ↓	Factors→						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
I_A		/					
I_B	/						
I_C	/						
I_D		/	/				
I_E			/				
I_F			/				/
I_G			/				
I_H			/	/			
I_I							
I_J							
I_K				/			
I_L					/		
I_M							
I_N	/		/				
I_O			/				
I_P			/				
I_Q				/		/	
I_R		/	/		/		
I_S						/	
I_T	/						
I_U					/		
I_V							/
I_W							
I_X							/
I_Y		/				/	

The main cluster of responses, provided in Table 106, refers to the use of intended learning outcomes in connection to criterion referencing (Driscoll and Wood, 2007, p.4). This is an approach cited by Alexander et al. (2008, p.309) as good practice for EAP courses where there is a need for a transparent linkage with assessment criteria, so that students know what is expected of them. Illustrative examples of comments which relate to the factors presented in Table 106 are listed below in Table 107.

Table 107 Examples of factors identified in BQ7

Factor number	Description of factors identified in interviews	Examples drawn from interviews (interviews coded using letters in brackets e.g. '(A)' denotes 'Interview A')		
1	Consider/based on existing assessments	B) I think a lot of it is based on what I've seen and reusing types of questions and ideas that are out there.	(C) I think we've never been brave enough at the institution I've worked at to start from scratch and I really think what the outcomes we want are. We're constantly fiddling around the edges with existing assessment methods.	(N) The problem was the tests will always upper intermediate regardless whether you had pre-intermediate students, so one thing was it had to be quite similar to previous tests.
2	Work as a pair or a team/use experience of others	(D) Every test we design, we always work at least as a pair, usually a group of three.	(R) I would use experience from my colleagues and go around and I'll ask them what they thought.	(Y) I think it would be really a team effort.
3	Link to intended learning outcomes	(D)I think we needed to be aware of what we're actually trying to get out of the student, what the learning outcome was and therefore, who does the question actually assess one of the learning outcomes.	(G)I mean, I'm actually choosing the particular question to test the particular aspect, you know?	(R)So you would be looking very much at saying, "Okay, what have you actually learned in this session? We're going to test you on what we think you should know."
4	Reference to question types	(H) I would ask a lot more, kind of in-depth questions to just kind of see that they understand kind of the whole holistic understanding of it.	(K) I try to have a development in the question, so we start with fairly discrete item questions about facts and figures in the text and we move then through questions which at time require more cognitive processing	(Q) Starting from very guided questions, to more sort of open questions, you know
5	Piloting/Trialling	(L) And then, if you can look at the items and see how they work, whether they are too easy, too difficult, or about right. You can also use an anchor test at the same time and compare your results with the anchor test.	(R) Actually test them out and see if they work. This is almost like a mini test	(U) We are setting in process a trialling process. And obviously once they've gone to that trialling process, we'll have more data and we can start eliminating them
6	Depends on level teaching	(Q) Again, it would depend on the level of the students that you were testing	(S) Certainly, making sure the questions were pitched at the right levels	(Y) So, it's no good sort of saying, "Well, I want to test students to see where they kind of C1." And then giving them a question which you know you're really going to get them up to about a B2, B1, sort of level.
7	Topic relevance/subject area	(F) the subject being covered in the lesson and which area I wanted to assess	(V) So I think the most difficult thing there is deciding on a question really and deciding on a topic	(X) I try to focus the question on something that is relevant to what they're doing.

4.8.4 Results of content analysis for items linked to views on/experience of validity and reliability in EAP testing and assessment

4.8.4.1 Major factors identified from interview Question BQ8 & BQ9

The findings which follow, drawn from items BQ8 and BQ9 and presented, in Table 108 below, are connected to the topic of validity and reliability in EAP testing and assessment. This section of results involves the conflation of responses to two separate questions given that responses overlapped due to confusion associated with the difference between validity and relaibility. Given that construct validity (Messick, 1989) can subsume both concepts, this has been deemed appropriate.

Table 108 Factors identified in BQ8 & BQ9 – interviewee views on reliability and validity in EAP testing and assessment

Interviews ↓	Factors→					
	(1)	(2)	(3)	(4)	(5)	(6)
I_A					/	
I_B	/					
I_C	/					
I_D	/	/				/
I_E	/			/	/	
I_F						
I_G	/		/			
I_H				/		
I_I				/		
I_J				/		
I_K	/			/		
I_L					/	
I_M					/	
I_N						
I_O			/	/		/
I_P				/		/
I_Q	/				/	
I_R	/		/			
I_S			/			/
I_T			/		/	
I_U	/	/	/			
I_V		/				
I_W						/
I_X	/					
I_Y			/		/	

When evaluating the factors which arise from interviewee comments, summarised in Table 108 above and exemplified in Table 109 below, Factor 1 shows the highest frequency of related comments. Factor 1 is characterised by interviewee lack of confidence with validity and reliability in EAP testing and assessment. Other key clusters include: piloting as a key aspect of pre-testing (Bachman & Palmer, 1996, p.234), the tracking of students, a key measure for the validation of EAP courses, as encouraged by BALEAP (2013).

Table 109 Examples of factors identified in BQ8 & BQ9

Factor number	Description of factors identified in interviews	Examples drawn from interviews (interviews coded using letters in brackets e.g. '(A)' denotes 'Interview A')		
1	Lack of confidence or admission of uncertainty	(Q) Validity, I would find quite difficult in that, you know, I think that's a very scientific area.	(U)We have no way of knowing at all.	(C) I have to admit to my eternal shame that we don't do that.
2	Look at the range of marks	(D) I would obviously look at it as telling that you have a huge range of marks and that your students of all abilities seem to be able to get a wide variety of marks.	(U) It would need to produce a spread of results. … we would expect to see that kind of spread mirrored in the result.	(V) If perhaps you've got maybe 5 or 10 who got A's and a few fails and the majority got you know, C. You would feel that it was a reasonable test.
3	Piloting/trialling	(G) Well I suppose you could always sort of guinea pig it on somebody and sort of get someone else to look at it and say, "What you think?"	(T) You have to do it on several times and so you can see if it will produce the same results with the same kind of candidates.	(Y) I think the things that I would do, would be obviously, piloting, trying the test.
4	Student tracking	(E) I suppose in the sense that we say if you've got 55, you can progress to your degree and then tracking the students.	(J) Improved feedback from Australia. So, five years ago, six years ago, our, our university had got some pretty terrible feedback. And since then every year, the feedback has continued to improve.	(P) I think the most logical way would be to measure it against the long-term goals that the students have got. How useful it is to them at university.
5	Teamwork and other teachers' views	(A) Working very closely with our externals.	(L) The other thing one can do is check teacher's views I think.	(T) A good question. It's difficult, I mean I'd seek peer confirmation, I think.
6	Moderation	(D)We look at moderation in terms of marking.	(O) I would need two or three markers and then compare results and have inter-rater reliability on it	(W) That's the big issue that we've got at the moment here with you know double marking, moderating. That's something…especially with the overall components. Yeah, I think, this is all right.

4.8.4.2 Major factors identified from interview Question BQ10

The information presented in Table 110, compiled after considering question BQ10, relates to factors emerging from interviewees' responses on the topic of content validity or domain representation in EAP testing and assessment. The controversies (see Dudley-Evans and St John, 1998; Spack, 1988 and Hyland, 2006) associated with viewpoints on this matter are often

indicative of individual practitioner's approaches to EAP and in many cases seem to charaterise choices made with regard to good practice.

Table 110 Factors identified in BQ10 – interviewee views on defining domain in EAP testing and assessment

Interviews ↓	Factors→				
	(1)	(2)	(3)	(4)	(5)
I_A					
I_B					
I_C	/		/		
I_D					
I_E		/			
I_F	/				
I_G	/				
I_H	/	/	/		
I_I				/	
I_J				/	
I_K				/	
I_L					
I_M				/	/
I_N			/	/	
I_O					
I_P	/				
I_Q				/	/
I_R				/	/
I_S	/				
I_T				/	
I_U	/	/			
I_V	/			/	/
I_W	/				
I_X					
I_Y				/	

With regard to means of defining or reflecting the domain associated with testing and assessment in EAP, Factors 1 and 4 from BQ10, in Table 110 above, stood out as having the most related responses. These factors, as exemplified in Table 111 below, suggest that interviewees attach importance to generalizing to real life, the challenges of which are explained by Douglas (2000,

p.12) through choosing subject matter in tests connected to the academic field of students as well as working in collaboration with academics from the relevant academic fields.

Table 111 Examples of factors identified in BQ10

Factor number	Description of factors identified in interviews	Examples drawn from interviews (interviews coded using letters in brackets e.g. '(A)' denotes 'Interview A')		
1	Choose subjects related to course or general for EGAP	(C) I think, we tend to just choose from, in inverted commas, serious subjects which are related to academic studies in some way.	(G) I sort of focus on the context	(P) Again, ideally it would be based on what people are going on to study so in a perfect world, you would be able to divide students according to if not subject, kind of faculty or area
2	Journals/Books	(E) looking at journals and books in the field of the study of the students	(H) I would try to find you know, academic journals	(U) find something which is an academic text written, written within a law department or in a legal journal or something like that.
3	Newspapers	(C) I think we tend to dumb down by choosing newspaper articles when we should be looking for suitable academic text	(H) newspaper articles or whatever that actually fit you for example.	(N) We used to use a lot of newspaper articles, something like this Telegraph one here
4	Work with academics from the field	(I) the best way, to be able to work with the professionals, although it isn't always the case that lawyers know the best about what is the best language to use.	(R) I think many universities have EAP and then you have the rest of the university and there has to be a bit more you know, more interaction	(Q) doing the sort of coordinating with the departments that they're going on to, finding out what is needed in those departments, and feeding that back into both your course design and your testing design
5	Difficult to work with academics from the field	(M) Well ideally, you know, in a perfect world, it would be nice to work with the subject teacher	(Q) But that's in an ideal world; there's often quite difficulties over communication between EAP tutors or departments	(V) If you can go up to the Law department and work out something with them then it's, it's very helpful for everybody if they are happy to do that, you know if you can get sort of communicate with the department.

4.8.5 Results of content analysis for items linked to practices involving analysis and interpretation of the results and scores of EAP tests and assessments

4.8.5.1 Major factors identified from interview Question BQ11

The findings from this next line of enquiry, drawn from responses to BQ11 and shown below in Table 112, reveal interviewee views on the topic of analysis and interpretation of EAP tests and assessments. It was intended that this area would provide information connected to practices which interviewees employ when they seek to give grades to students' work and to draw inferences from material relating to students' abilities. This area also relates to arguments connected to the need to use educational objectives (Bloom, 1956) and outcomes-based approaches to asssement Driscoll and Wood (2007).

Table 112 Factors identified in BQ11 – interviewee views on analysis and interpretation of EAP testing and assessment

Interviews ↓	Factors→			
	(1)	(2)	(3)	(4)
I_A	/			
I_B	/			
I_C		/		
I_D	/			
I_E		/		
I_F				
I_G	/	/		
I_H	/			
I_I			/	
I_J				
I_K				/
I_L	/	/		
I_M	/			
I_N			/	
I_O				/
I_P				
I_Q				
I_R	/			
I_S	/	/	/	
I_T	/			/
I_U				
I_V	/			
I_W	/		/	/
I_X				/
I_Y				

The main pattern of responses to BQ11 indicates, through Factor 1, that large groupings of interviewees consider the use of grading criteria to play a key role in the assessment of EAP. Examples which illustrate the factors identified in Table 112 above are presented below in Table 113.

Table 113 Examples of factors identified in BQ11

Factor number	Description of factors identified in interviews	Examples drawn from interviews (interviews coded using letters in brackets e.g. '(A)' denotes 'Interview A')		
1	Criteria are needed	(B) we've got set of criteria so, I'm working against institutional criteria that have been pre-set.	(H) I think if they met then that criteria, that final criteria, you know, I guess if they could cope at that next level.	(L) suppose they developed from the IELTS approach where there are band descriptors… there are descriptors
2	Reference to use of percentage marks	(E) I don't know if this is right…I mean, in our course, it's been, the setup that 55% for internal purposes is considered to equivalent to IELTS 6.5.	(G) I do tend to do percentages because I can compare people and strengths and choose percentages for individual skills.	(S) it's really important to know that actually, 50% says that they have done this, this and this but haven't quite managed to do that.
3	Moderation/compare with others	(I) I would hope to have the help of at least one other marker	(N) So standardisation means that, for example, we look through a whole lot of tests.. and you have to try to agree a level	(W) And then we do blind marking anonymised, yeah, blind marking. And then we do the sample moderation
4	Unscientific/subjective	(K) So I think there is a great deal of subjective judgment comes in whatever we say about the test results.	(O)I don't have a huge conceptual framework for that… I can't answer that really well	(T) I mean I think that you know my idea about testing is it's…it's fairly…it's not an exact science.

4.8.5.2 Major factors identified from interview Question BQ12

The factors identified in Table 114, drawn from responses provided to item BQ12, relate to interviewees' opinions on the accuracy of their interpretations in the process of EAP testing and assessment. It was hoped that this item would help to identify whether repondents have any concerns in this area. This question also links with the concept of construct validity, given the need for inferences drawn from tests to be trustworthy (Gipps, 1994; Popham, 2001; Stiggins, 1991)

Table 114 Factors identified in BQ12 – interviewees' views on accuracy of interpretation in EAP testing and assessment

Interviews ↓	Factors→				
	(1)	(2)	(3)	(4)	(5)
I_A	/	/			
I_B			/		
I_C		/			
I_D				/	
I_E				/	
I_F					/
I_G		/			/
I_H					
I_R				/	
I_J	/		/	/	
I_K					
I_L		/			
I_M			/		/
I_N				/	
I_O					
I_P					
I_Q					
I_R				/	
I_S	/				
I_T					/
I_U		/	/		
I_V				/	
I_W					
I_X					
I_Y		/			

Table 114 shows that the two most frequently cited factors are Factor 2 and Factor 4. Factor 2 involves interviewees' admission that the process of interpreting EAP test results is unscientific or inaccurate, raising concerns similar to those suggested by Coniam (2009), regarding in-house EAP testing. Factor 4 makes reference to the use of moderation as a means of enhancing accuracy of interpretation. Factors 1-5, as identified in Table 114, are illustrated with examples in Table 115 below.

Table 115 Examples of factors identified in BQ12

Factor number	Description of factors identified in interviews	Examples drawn from interviews (interviews coded using letters in brackets e.g. '(A)' denotes 'Interview A')		
1	Some interpretation is required	(A) There's always some interpretation that somebody has.	(K) I guess it comes down to my understanding of education as an inherently subjective, interpersonal, and intrapersonal process.	(S) I think it's just, over the years, it's past experience. We've done it before. It becomes easier in the years as you should go along, being able to interpret the results of something.
2	Not scientific	(C) Unfortunately it is not done in a very formal way. It is purely anecdotal at the moment.	(G) I don't think that they are. I don't feel confident they're completely accurate	(L) Well I don't think you do actually. I mean, I don't think there is any such thing as, you know, a true score because whatever you've got in terms of the score, there will always be something, some variants that you cannot account for.
3	Use of Criteria	(B) I put a lot of trust in them (criteria	(J) you just have to have very specific band to scriptures to ensure that everyone's working at the same level, I think.	(M)What assessment criteria are there? And in terms of accurately reflecting what the student actually is. I mean, that's the problem for all grading isn't it?
4	Moderation	(E) we are sort of going in the direction now of having to do more moderation	(N) Basically you would ask the different people to mark them. You come to a session where you talk them through….	(R) you do moderation, you actually look and you see you know, is someone marking higher than others?
5	Consider other results or patterns	(G) I do do this comparison and I do look back at the tests that are way out of sync and think, well, have I done something wrong.	(F) You'd have to look at other tests that they've done and I think it's always important to	(T) I am reassured by patterns of consistency with myself which I'm…you know and with others. And disturbed by the opposite which does sometimes happen

4.8.5.3 Major factors identified from interview Question BQ13

The data collected from this section of the interview, selected from responses to BQ13 and arranged below in Table 116, deals with the subject of assisting stakeholders in the act of understanding results from EAP tests and assessments. The aim of this question was to ascertain the extent to which interviewees are actively involved in providing advice and feedback to different stakeholders so that test results can be made sufficiently meaningful. In this way, the

question involves Assessment Literacy in connection with assessment for learning, as described by Knight (1995) and Price et al. (2012).

Table 116 Factors identified in BQ13 – interviewee views on helping stakeholders to understand grades

Interviews ↓	Factors→		
	(1)	(2)	(3)
I_A			
I_B	/		
I_C			
I_D		/	
I_E			/
I_F		/	
I_G		/	
I_H			
I_I			
I_J	/		/
I_K			
I_L	/		/
I_M	/	/	
I_N		/	
I_O			
I_P		/	
I_Q			/
I_R			
I_S		/	
I_T			
I_U			
I_V			
I_W			/
I_X		/	
I_Y	/	/	

The main area of commonality which emerged from BQ13, as shown in Table 116, was connected with the provision of detailed formative feedback, as noted by comments which cluster around Factor 2. Other common threads reveal a focus on interpreting final scores and the use or creation of IELTS equivalences. Respondent comments which exemplify the factors identified in Table 116 are presented below in Table 117.

Table 117 Examples of factors identified in BQ13

Factor number	Description of factors identified in interviews	Examples drawn from interviews (interviews coded using letters in brackets e.g. '(A)' denotes 'Interview A')		
1	Looking at final scores	(B) A lot of time they just look at the final score and that final score tells them if they're good enough or not good enough.	(J) What we do is we, we have a pass grade and then we say what the merit would be and what an outstanding student would be.	(Y) I think what we find is that the department, I mean…how direct can I really be? I mean with department, I think they offer one kind of quick and dirty answer don't they really?
2	Detailed feedback (formative)	(M) much more emphasis is placed on the feedback basically rather than the final grade	(G) students will get feedback on their assessments and their tests and so they will understand them as best we can explain, you know, in terms of formative feedback	(S) Well, I, always feedback on marks and that will always involve things that they have achieved
3	IELTS equivalence	(J) But it also, we try to equate all of those levels as well to an IELTS score. So that the university that the student go to will have an understanding of the level	(L) The university will understand that is the same as on IELTS 6.5. It does not matter how often you tell them actually, it is not the same because we had used different approaches to measuring skills	(W) we end up saying it's equivalent to, you know, IELTS 7.5 or whatever, you know and I want to stop doing that because, you know, it isn't.

4.8.6 Results of content analysis for items linked to the influence of research and other resources for the purpose of EAP testing and assessment

4.8.6.1 Major factors identified from interview Question BQ14

Factors emerging from responses to question BQ14 are described in Table 118 below. This item was designed to elicit responses on the extent to which interviewees feel that their practice has been influenced by research associated with EAP assessment and testing. This section hoped to identify how proactive or engaged interviewees were with research in the field of EAP testing and assessment and whether interviewees were actively trying to develop their own Assessment Literacy in this way. This links directly to the research question which refers to research-led practice. Such activity forms part of the process which Stiggins (1995) describes as approaching

assessment with the full knowledge of what is being assessed. It also aligns with features of practice pertaining to educators which Menter (2010) describes as transformative.

Table 118 Factors identified in BQ14 – interviewees' comments on the extent to which they are influenced by research related to EAP testing and assessment

Interviews ↓	Factors→			
	(1)	(2)	(3)	(4)
I_A	/			
I_B				
I_C		/		
I_D		/		
I_E		/		
I_F		/		
I_G		/	/	
I_H		/		
I_I		/		/
I_J		/		/
I_K		/		/
I_L	/			/
I_M	/			
I_N		/		
I_O		/		
I_P	/			
I_Q		/		
I_R		/		
I_S	/			/
I_T		/		
I_U				
I_V	/			
I_W	/		/	
I_X			/	
I_Y	/			/

Whilst factors 1, 3 and 4, which emerged from responses to BQ14, show that some interviewees do report being influenced by research relevant to EAP and assessment, Factor 2 shows the highest frequency of response showing a large number of interviewees who claim not to have been influenced by research in EAP assessment practice. It could be considered that this behaviour does not facilitate an expanded Assessment Literacy, which practitioners require in order to make the

right decisions for language learners and institutions (Taylor, 2009). Table 119 provides examples of factors which were noted in response to BQ14.

Table 119 Examples of factors identified in BQ14

Factor number	Description of factors identified in interviews	Examples drawn from interviews (interviews coded using letters in brackets e.g. '(A)' denotes 'Interview A')		
1	Yes I'm influenced by research	(L)Well, quite a lot really, I have a range of things on my shelf	(P) Just at the moment a lot. Just because I'm in the middle of doing this literature review. You have asked me at the right time.	(W) Yeah, I mean more recent…that, I mean because that was a long time ago and things have got a bit rusty on things as you can probably tell
2	No, not influenced much	(D)Probably not huge amount, to be very honest.	(H) Erm, I don't really think it's influenced	(J) I think directly, not at all. It would be the honest answer
3	BALEAP or IATEFL influence	(G) I've been looking at the BALEAP teaching competencies and they also refer to assessment	(W) But I like to read you know, when I get time, journal articles too, the testing SIG for IATEFL	(X) I do keep an eye on what, for example the IATEFL organisation are looking at.
4	Reference to literature consulted	(K) I've got a small number of books on testing and so on.	(S) All I know is that those ideas have come from the literature that I've read.	(Y) I think my main influences are really in terms of the kind philosophical dimension of language testing really.

4.8.6.2 Major factors identified from interview Question BQ15

The results presented below in Table 120, drawn from responses to BQ15, reveal responses associated with particular books which interviewees were able to refer to, as another measure of familiarity with key research in the fields of EAP and assessment.

Table 120 Factors identified in BQ15 – interviewees' comments on the use of particular books or journals

Interviews ↓	Factors→		
	(1)	(2)	(3)
I_A			
I_B	/		
I_C			
I_D	/		
I_E	/		
I_F	/		
I_G	/		
I_H	/		
I_I			
I_J		/	
I_K			
I_L		/	
I_M			
I_N			/
I_O		/	
I_P		/	/
I_Q	/		
I_R	/		
I_S			/
I_T	/		
I_U	/		
I_V			/
I_W			
I_X	/		
I_Y		/	

The most prevalent factor, which was identified through BQ15, was Factor 1 which showed that 11 interviewees do not claim to consult any particular books or journals. Factors 2 and 3 show interviewees' reference to specific titles of books or journals, in addition to their use of books and journals more generally. Table 121 shows examples of factors arising from BQ15.

Table 121 Examples of factors identified in BQ15

Factor number	Description of factors identified in interviews	Examples drawn from interviews (interviews coded using letters in brackets e.g. '(A)' denotes 'Interview A')		
1	I don't consult anything	(F) No, I haven't yet.	(Q) Not for testing, no, no	(T) There's nothing in mind for me. I have to sort of start fresh …
2	Makes reference to specific books or journals	(J) I don't know if the name John Swales rings any bells but…he used to work for a university. He used to prepare everyone for the academic English test.	(O)A text on my course is – I can't remember the name of it. It's Hughes 2003?	(Y)I mean I think the Cyril Weir 2005 book, language testing and validation. I mean, I mean that you know for me was quite an influential book
3	Refers generally to using books, articles or journals	(N) Well, I don't know about specific. I mean I would be going on to look at online journals	(P)I wouldn't say that there is a particular journal that I'd go to. I search the keywords.	(V) I've got materials but it's more like sort of again you know because you know because the students are from different academic backgrounds.

4.8.6.3 Major factors identified from interview Question BQ16

The patterns which have emerged from this area of the results section, taken from interviewee responses to BQ16 and listed in Table 122, are concerned with identifying any additional resources which interviewees believed could assist in the process of EAP assessment and the development of Assessment Literacy. It was envisaged that this question that this would highlight further any areas where interviewees felt the need for specific support.

Table 122 Factors identified in BQ16 – interviewees' comments on their need for additional resources for testing and assessment of EAP

Interviews ↓	Factors→			
	(1)	(2)	(3)	(4)
I_A	/			
I_B	/			
I_C				
I_D		/		
I_E	/			
I_F	/			
I_G		/		/
I_H				
I_I			/	
I_J				
I_K				/
I_L				
I_M				
I_N				
I_O				
I_P			/	
I_Q				
I_R	/			
I_S		/		
I_T				
I_U		/		
I_V				
I_W				
I_X			/	
I_Y				/

The highest cluster of responses indicates that five interviewees referred to their need for additional training associated with EAP assessment. A further four interviewees mentioned a desire to gain access to different versions of EAP tests or questions. A series of examples which illustrate the patterns of response to BQ16 are provided in Table 123.

Table 123 Examples of factors identified in BQ16

Factor number	Description of factors identified in interviews	Examples drawn from interviews (interviews coded using letters in brackets e.g. '(A)' denotes 'Interview A')		
1	Training	(A) I know there's training out there	(F) more chance to do training in those kinds of areas,	(R) I think it would be very useful for us to have had some kind of training in what the criteria actually means.
2	Access to different versions of EAP tests or question types	(D) What I'd like is different versions of tests, or different ideas of what type of questions can be asked	(G) as a practitioner, you want to be able to pick up a few samples of EAP tests and they don't…it seem to me…and I know I haven't looked hard enough, but there don't seem to be readily available samples	(U) I think, I think I would like to have a text or a course which…provided with clear guidelines, clear kind of instructions helping you with question types and that kind of thing.
3	More time	(I) More time	(P) Time and money.	(X) I'm frankly busy enough with everyday teaching to do
4	Teamwork on a testing project/focus	(G) I'd say maybe a more formal structured project where it's agreed that because a lot of people are going to work on tests and then, you know, a procedure, I think that would help.	(K) I wonder what pressure it would take for us to actually sit down and work out together what we mean when we talk about assessment… Because I think if we seriously did that, I think it would be a tremendously productive process.	(Y) I think the only thing that I would welcome I think would be more people being involved in testing. And then it goes back to that community and sort of the collaboration thing.

4.8.7 Results of content analysis for items linked to the consideration of ethics in EAP testing and assessment

4.8.7.1 Major factors identified from interview Question BQ17

The findings from this next section, collected from BQ17, focus on interviewees' concerns connected with EAP testing and assessment. It was hoped that this item would allow respondents to identify any issues of concern or ethical worries that they have encountered. This might include issues connected to the social implications of testing (McNamara and Roever, 2006) and the associated power relations (Shohamy, 2001).

Table 124 Factors identified in BQ17 – interviewees' concerns relating to EAP testing and assessment

Interviews ↓	Factors→						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
I_A							
I_B							
I_C							
I_D	/	/					
I_E							
I_F	/						
I_G							
I_H							
I_I			/	/			
I_J	/			/	/		
I_K				/			/
I_L		/		/			
I_M				/	/		
I_N			/	/	/		
I_O	/			/			
I_P		/		/			/
I_Q					/		
I_R	/			/			
I_S						/	
I_T				/			
I_U						/	
I_V				/		/	
I_W				/			
I_X							
I_Y			/				/

The most noticeable grouping of responses to emerge from BQ17 is based on twelve interviewee comments which express concerns relating to IELTS and its perceived lack of reliability. The next main factor, cited by five interviewees, also refers to concerns associated with reliability in EAP testing and assessment. Examples illustrating the factors from BQ17 are listed in Table 125.

Table 125 Examples of factors identified in BQ17

Factor number	Description of factors identified in interviews	Examples drawn from interviews (interviews coded using letters in brackets e.g. '(A)' denotes 'Interview A')		
1	Validity/Reliability	(D) there is an ethical point on whether or not marks are actually a true reflection of students' ability	(F) And you wonder if that's really sort of represents to their ability	(R) I'm not so sure that the test marks, very often do show what the student can do. I don't think they're always that accurate.
2	Pressure and Power	(D) there's a pressure to make sure that students get through	(L) There's so much riding on it and the decisions which are made based on test scores and it can affect people's lives. Also the impact of IELTS and other similar tests is getting more and more powerful.	(P) Because there's a lot more riding on it, both in terms of people's future careers and money as well.
3	Washback, pervasive nature of testing	(I)it is all pervasive, it's in every country that you go to and people are training to take tests like IELTS and answer questions rather than learning the language. And the results of tests like IELTS seem to be very variable.	(N) Somewhere like China, you just see the sheer volume; you don't get the same idea in Britain at all. IELTS is the kind of machine which it has thousands of students doing it every month… So that is an impact of the exams I find very problematic in terms of trying to make someone really develop their skills	(Y) What you then get is the issue of washback. Because the things that you're testing the students on are not necessarily the things, I mean they might be things that are covered on the course
4	IELTS variable / unreliable	(K) I've got real concerns about the larger-scale kind of testing, especially IELTS.	(M) I think there is a well-documented problems and performance on IELTS	(T) I'm not sure about IELTS. I've taught IELTS and I think IELTS is always kind of obeys its own laws.
5	A general or specific focus on EAP domain	(J) So I think that the teaching is often too general.	(M) assessing them not on specific EAP genres, let's say, and kind of possibly disadvantaging them	(Q) How EAP actually relates to the real world that follows on from it. I think that is an issue
6	EAP/ Pre-sessional courses are positive	(S) It's totally essential and I think the students appreciate it.	(U) I think it should go on. And I kind of glad that it does. I think it has a useful function.	(V) But I do feel that you know the EAP courses are very helpful for them
7	Government UKBA	(P) the Home Office needs an IELTS or asks for an IELTS score. So, people understandably trust it.	(K) if the home office gave us the opportunity to design and validate our own assessments as we do with our own groups, we could go with something that would give us more insights into the student.	(Y) I think with sort the stringent, sort of UKBA requirements in place. There can't really be any doubt about that. … You have to; you have to be able to be absolutely certain.

4.8.7.2 Major factors identified from interview Question BQ18

Responses provided relating to the next area of the interview, compiled from responses to BQ18 and shown in Table 126, involve interviewees' views on whether EAP testing and assessment requires different considerations in comparison to other forms of language assessment. The aim of this question was again to identify any particular features connected to EAP assessment which interviewees thought might warrant specific consideration.

Table 126 Factors identified in BQ18 – interviewees' comments regarding the extent to which EAP testing and assessment requires different consideration to that of testing general English

Interviews ↓	Factors →
	(1)
I_A	/
I_B	
I_C	/
I_D	/
I_E	/
I_F	
I_G	/
I_H	
I_I	
I_J	
I_K	
I_L	
I_M	
I_N	
I_O	/
I_P	
I_Q	/
I_R	/
I_S	/
I_T	/
I_U	/
I_V	/
I_W	/
I_X	/
I_Y	

The results of this section showed, through Factor 1, that 14 interviewees cited the need for EAP to be given special consideration in a manner distinct to other forms of language assessment, given that EAP represents a different genre of English or requires the development of a different set of skills associated with academic practice or different academic domains. Table 127 displays examples of responses clustering in Factor 1.

Table 127 Examples of factors identified in BQ18

Factor number	Description of factors identified in interviews	Examples drawn from interviews (interviews coded using letters in brackets e.g. '(A)' denotes 'Interview A')		
1	Due to different genre and skills	(C)Yeah enormously.. I think...we've got to think what does a student do as part of the...undergraduate programme linguistically?	(S) Yes, you have got to really focus on academic things which if you're doing general language testing you wouldn't be concerned with	(V) it's the academic research skills that they are learning. And hopefully you know showcasing in their work. Whereas in straight language testing it's just the language you know

4.8.7.3 Major factors identified from interview Question BQ19

The data collected from this section of the interview, given in response to item BQ19 and listed in Table 128 below, allowed interviewees to provide any additional comments that they may have wished to add. This question performed a similar function to the initial open question at the start of the interview.

Table 128 Factors identified in BQ19 – interviewees additional final comments

Interviews ↓	Factors→ (1)
I_A	/
I_B	/
I_C	/
I_D	
I_E	
I_F	/
I_G	
I_H	
I_I	
I_J	
I_K	/
I_L	/
I_M	/
I_N	
I_O	
I_P	/
I_Q	/
I_R	
I_S	/
I_T	
I_U	/
I_V	
I_W	/
I_X	
I_Y	

The only common factor observed by the researcher relating to the final question was that the interviewees mainly did not feel the need to add additional comments. Examples of responses which grouped in this factor are provided below in Table 129.

Table 129 Examples of factors identified in BQ19

Factor number	Description of factors identified in interviews	Examples drawn from interviews (interviews coded using letters in brackets e.g. '(A)' denotes 'Interview A')		
1	None	(L)No, not really that's all thanks.	(S)There's nothing to say	(W)No, not really

4.8.8 Section conclusion

The findings presented in this section have revealed clusters of responses which have emerged from an interview protocol. The interview framework has been designed to elicit views and described behaviours from EAP tutors which can be considered with regard to the extent that they reflect language testing research and good practice associated with Assessment Literacy. This information will be of particular use when analysed in Chapter 5, in conjunction with the findings which have been gathered from the questionnaire tool.

4.9 Chapter conclusion

In this Chapter, the information collected through the questionnaire and interview research tools has been presented in a manner which will allow the data to be reviewed, considered and analysed further in Chapter 5. The findings have been organized according to a series of sets which relate to the research methods that have been employed. These sets are summarized below:

- Set A1: Features of questionnaire respondent population and identification of a relevant population for the study
- Set A2: Descriptive statistics for individual questionnaire items:
- Set A3: Composite results of questionnaire Likert scales and associated statistics
- Set A4: Content analysis of open questionnaire
- Set B1: Interview data analysis- sample population description
- Set B2: Interview interactional profiles
- Set B3: Content analysis of interview question responses

In order to display the findings in the context of Assessment Literacy, as required by the research questions, results within each set have, where possible, been grouped according to the aspect of Assessment Literacy which they refer to. This has been based on key areas of Assessment Literacy in EAP which were identified for investigation at the end of the Literature Review.

CHAPTER 5 ANALYSIS AND DISCUSSION

5.1 Chapter introduction

In this next section of the thesis, the Findings, as presented in Chapter 4, will be analysed and discussed. During this process the results of research will be discussed in relation to key literature and theoretical frameworks which emerged from the Literature Review in Chapter 2. The implications of the analysed findings will then be reviewed in relation to the research questions in order to identify unique inferences which can be drawn.

The Analysis and Discussion will be structured in a manner which deals with each of the different blocks of data, presented in the Findings which have been collected through the main research tools in the form of the questionnaire and the interviews. However, in order to present the analysis in a manner which affords the requisite prominence to the extrapolations which the researcher considers to be of most importance, the order in which interpretations are presented will differ to the linear manner in which the Findings from each of the research tools was presented in Chapter 4.

Initially, a discussion will be presented relating to the analysis of data from both Set A1 and Set B1 connected to the sample populations for both the questionnaire and interviews. The discussions will then move on to discuss the analyses which have been based on Set A3, which explores the implications of findings drawn from the composite results of questionnaire Likert scales. These clustered results, which have been compiled in conjunction with Factor Analysis and reliability testing can be considered more powerful than individual items. Inferences which emerge from these results will then be considered in conjunction with additional interpretations which arise from the individual DVAS items, as displayed in the results in Set A2, some of which did not

contribute to the composite scales. The analysis will then proceed to consider the open questionnaire responses from Set A4, before examining the results of the content analysis of the interviews conducted, the findings for which are presented in Set B3. These interpretations will then be considered in line with the interactional profiles provided in Set B2.

5.2 Discussion and analysis of sets Set A1 and Set B1 related to population samples

5.2.1 Section introduction

Sampling is of crucial importance to the analysis stage of research, as it is usually not possible to collect data from all members of a relevant population. Consequently, choices need to be made regarding how members of a population can be selected for participation, in a way which minimises limits on the conclusions which can be drawn from research and the extent to which they may be considered generalisable (Gorard, 2001, p.11; Miles & Huberman, 1994, p.27).

During the use of the main two research tools in this study, care was taken to select a representative sample of EAP teachers so that the results could be considered to provide an accurate representation of practitioners in the field of EAP. What follows in this section is an analysis and discussion of the data in Sets A1 and B1 which clarifies the approach used to sample a relevant population of EAP tutors who were involved in the practice of assessment, in order to align with the nature of the research questions.

5.2.2 Discussion of population sample used for further analysis

Sections 3.4.5 and 3.4.9 show the researcher's approach to attracting participants to take part in the study, through the medium of online professional networks associated with EAP located both within the UK and overseas. Through this method, an initial population of 187 respondents completed the questionnaire in full. Sections 4.2 and 4.6 of the Findings in Chapter 4 present

features of Sets A1 and B1 showing the total population of questionnaire respondents and interviewees along with the manner in which these samples were reviewed and, where necessary, filtered or stratified (See 4.6.2), in order to ensure that only a balanced and representative sample was considered at subsequent stages of presenting findings and ensuing analysis.

With regard to the questionnaire respondent population, consideration has been given to the characteristics of questionnaire respondents, in particular, the balance of age, gender, nationalities and locations of practice in conjunction with respondents' acknowledgement of association with the field of EAP and institutional affiliation. As regards the interview population, a similar pattern of features was engineered in the sample of interview respondents through soliciting participation from interviewees, via email contact details provided at the questionnaire stage. Through this process, a comparable demographic pattern could be replicated to that of the questionnaire. Consequently, based on the results of Sections 4.2 and 4.6, relating to Sets A1 and B1 and population-related data, it is deemed by the researcher that the sample features in both the questionnaire and interviewee populations are appropriately representative of each other and of the sector. This therefore suggests that they are also suitable for the purposes of generalisation beyond the populations within this study.

With regard to sample sizes, the researcher recognises the key difference in volume of cases associated with each of the two main data collection methods. Nevertheless, the view is also held that the nature of the samples and their respective quantitative and qualitative orientation remain complementary and concomitant with the Pragmatist paradigm adopted by this study, as described in Table 24 on page 21. The researcher also supports the fact that the difference in sample sizes reflects the relative merits which can be obtained from a mixed methods approach. It is believed that the mixed methods approach enhances the validity of inferences drawn from this study. The

researcher shares the view upheld by Duff (in Chalhoub-Deville et al., 2006, p.66), who advises that the polarisation of qualitative and quantitative methods is often exaggerated. In addition, opposing concerns, such as those expressed by Adler and Adler (1994, p.377-92), regarding the limitations of qualitative research (in Nesfield-Cookson, 1987 in Cohen and Manion, 2007, p.17) and the mechanistic approach of a solely quantitative study, can also be considered to highlight the benefits of combining both approaches, as detailed in Chapter 3 of this thesis.

For the questionnaire, an initial count of 187 respondents submitted each section of the tool, which was subsequently filtered to a sample of 158 respondents, with a lower number of sub-categories, for consideration in the main study. In accordance with guidance provided by Gorard (2001, p.17) and Denscombe (2007, p28), who describe key features of sample sizes which can be considered larger in scale and sufficiently valid or reliable, the questionnaire sample can be seen to conform satisfactorily. The overall questionnaire sample and subsamples within the questionnaire, which relate to categories used for gender, nationality and location of practice, are also big enough to be considered significant and acceptable for statistical analysis.

With regard to the sample size for the interviews, although just 25 interviews were conducted, the aim of the interview stage was to collect a smaller yet still purposeful sample of data which allows an in-depth understanding of the phenomenon being explored (Heigham & Croker, 2009, p.149), namely EAP teacher Assessment Literacy. Given that the population of interviewees approached has been stratified to comply with the features of respondents to the questionnaire by identifying interviewees from the random sampling during the questionnaire stage, the researcher deems that the level of sampling has been sufficiently varied and rich. It is thought that this will allow the elicitation of inferences from the interviews which will complement those gleaned from the questionnaire.

5.2.3 Section conclusion

In this section, the rationale and procedure for sampling a relevant number of EAP teachers from the wider target population of individuals for the questionnaire and interviews has been provided. The steps which have been taken and the approach which has been adopted have explained how the researcher has endeavoured to gain access to a sample of data that is both relevant and generalizable to a larger representative group. This has been a crucial stage of the research, given the need to show that that the people from whom data has been collected accurately reflect the target group as identified in the thesis research questions. In many ways, this stage of the research can be seen to represent the foundation for the rest of the study.

5.3 Discussion and analysis of Set A3 - Composite results of questionnaire Likert scales

5.3.1 Section introduction

The results of the composite Likert scales sum together related scales from the questionnaire which have been identified as valid and reliable through Factor Analysis, reliability statistics and researcher judgement. Consequently, it can be argued that inferences drawn from these findings are more powerful than those drawn from individual items, given that they represent clusters of opinion on related topics from groups of respondent who make similar choices.

The areas of questioning which were used within the different sections of the questionnaire, and which were intended to represent the theoretical framework for Assessment Literacy in EAP, were inspired and steered by principles such as those suggested by Popham, (2001) and Stiggins (1995) and identified during the Literature Review for this thesis.

The results from each composite Likert scale will now be evaluated in order to consider the implications. In order to focus on the meaning, the factor titles as identified in Table 65 on 270 will be used. These factors will also be presented according to the structure used to organise the results as presented in the Findings chapter, along with the key areas of Assessment Literacy under exploration, which were identified for exploration in the Literature Review.

5.3.2 Discussion of factors relating to training, skills, strengths and weaknesses in EAP testing and assessment

5.3.2.1 Analysis of 'Respondents' interest and confidence in EAP language testing and assessment'

The composite results relating to this factor (drawn from AQ13 D,E,F and G) represent a sub factor drawn from the section of the interview intended to collect respondents' views on their interest, skills and training related to EAP testing and assessment.

This particular cluster focuses on questions associated with EAP teachers' engagement with and confidence in EAP assessment. In particular the items ask respondents to comment on the regularity with which they consult books on language testing, the extent to which respondents feel EAP assessment represents a key skill for them, whether they find support material user friendly and perceptions regarding the need for additional training.

As the results displayed in the Findings chapter demonstrate, only just over one third of respondents agreed or strongly agreed in response to the items in this group. This also left just under 40% of respondents who disagreed. When those respondents who were unable to either agree or disagree are taken into account, it becomes clear that over 66% of respondents did not

reply in a manner which indicated a high level of confidence and interest in EAP assessment and testing. These findings have key implications for the research questions given that they act as a measure regarding practitioner views of their own proactive engagement and capability in EAP assessment practices.

The aspects of Assessment Literacy, as defined by researchers such as Popham (2001) and Stiggins (1995), require a proactive and informed approach to assessment due to the implications of poor assessment for stakeholders involved. As a result, any indication of less than optimal interest and low confidence amongst practitioners identifies key hurdles to the promotion of the principles of Assessment Literacy across the EAP teaching community. Concerns are also raised relating to responsibility for stakeholders and the outcomes of tests, as warned by critical assessment specialists (McNamara, 2000; Shohamy, 2001; Stobart, 2008).

The implications of these results will be considered further in conjunction with reference to respondents' views on their EAP assessment skill set which will be discussed later during this section of the analysis.

5.3.2.2 Analysis of 'Respondents' perceptions regarding complexity of EAP language testing and assessment'

The items in this factor (AQ13 H, I and J) are intended to act as a measure of the extent to which respondents consider various aspects of EAP assessment to be complex. The results show that just over one third of respondents agree or strongly agree thus recognising the complexity of EAP assessment in the situations presented. In contrast, nearly 40% disagree or strongly disagree, inferring agreement that the cited aspects of EAP assessment were straightforward. 27% of respondents remain unable to agree or disagreed that the identified aspects of EAP assessment

were complex. In particular, the focus of the items is on the identification of the features of EAP in which students need to be assessed, the recognition of faulty test items and the relative complexity of language assessment compared to subject knowledge.

The interpretation of these results can be viewed from different perspectives, based on how the terms complex or straightforward are considered. In some circumstances, an acknowledgement of a task's straightforward nature could be related to a practitioner's level of skill. However, it should be noted that experts in the field of language testing acknowledge the inherent and enduring complexity associated with defining and operationalizing language constructs in assessments (Bachman & Palmer, 1996, p. 62; Fulcher & Davidson, 2007; p.7). As a result, the researcher has opted to consider responses indicating the straightforward nature of EAP assessment to demonstrate a relative underestimation of the challenge presented and therefore an indicator of deficit in Assessment Literacy.

With this standpoint in mind, it can be interpreted from the data that 45.1% of respondents agree or strongly agree that EAP assessment is complex in the manners described above, whilst 31.5% disagree or strongly disagree. This has left 23.4% of respondents who neither agreed nor disagreed that the identified aspects of EAP assessment were complex.

As it can be claimed that only just over one third of respondents acknowledged the complexity of EAP assessment, it can be inferred that further training or provision is required in order to explain in further depth the challenges which assessment of this nature presents. Through additional training, challenge of addressing the factors associated with Assessment Literacy in EAP can adequately be prepared for by more practitioners on a sustained basis in a manner which

recognises language testing as ‘a creative, organic, consensus driven, iterative process’(Davidson & Lynch, 2002; Fulcher and Davidson 2007, p.61).

5.3.2.3 Analysis of ‘Respondents’ skills in aspects of EAP assessment’

The items which are combined and related to this particular construct are drawn from AQ14 and relate to respondents’ self-evaluation of skills related to EAP assessment. These skill areas were selected for inclusion based on their relevance to both Assessment Literacy, EAP and the Test Design Framework (Fulcher, 2010, p.94), as discussed in Chapter 2.

The results of the composites show that 38.9% of respondents consider themselves to be either highly skilled or very highly skilled, although only 10.7% describe themselves as low skilled or very low skilled. Finally, nearly 50% of respondents rate their skill level as average in the areas identified.

Whilst it can be interpreted as positive that only just over 10% of respondents consider their skill level to be low or very low in level, it is noteworthy that approximately 50% position their ability as average. Based on this situation, it can be interpreted that although an average position may not be deemed to identify a critical development need, given the high stakes associated with EAP assessment (Alexander et al., 2008, p.307) a higher rating of skill from amongst the population would be desirable. A higher skills rating amongst respondents would help to ensure that the quality, validity and reliability of testing and the inferences which are drawn from test scores can be maintained at as high a level as possible. As a result, this suggests that interventions are necessary in order to raise EAP teachers’ level of Assessment Literacy in EAP.

5.3.3 Discussion of factors relating to approaches to EAP testing and assessment design

5.3.3.1 Analysis of 'Respondents' views on processes and procedures in EAP test design'

Items contributing to this factor are drawn from AQ16 (AFG) and involve respondents' views on processes and procedures in EAP test design, in particular the starting point for test development and the need for teamwork in line with the view of test development, as favoured by Davidson and Lynch (2002) and Fulcher and Davidson (2007, p.61).

The results show that a majority of 60% of respondents agreed or strongly agreed with certain statements. These statements indicate an understanding of the importance of starting the test design process at a point other than with writing test items, the significance of teamwork and the benefits to practitioners of gaining experience of test design over time. In contrast, just 13.5 % of practitioners disagreed with the significance of the factors mentioned above, leaving 26.6% of respondents who neither agreed nor disagreed.

This grouping of factors and the majority view expressed by respondents aligns with the experience of test developers who recognise the benefits of a team approach in the development of tests (Fulcher and Davidson 2007, p.61). The views expressed also acknowledge respondents' understanding of two crucial types of resource for the purposes of language testing in the form of adequate time and sufficient human resources (Bachman & Palmer, p.37).

Overall, the findings emanating from this area of the research can be considered to show that a large proportion of respondents understand the value of time investment, team work and planning in advance of question creation in the process of test and assessment testing. This being the case, there is also scope for providing additional support for those remaining respondents who make up approximately 40% of the total group and whose response did not correlate so noticeably positively with indicators associated with good Assessment Literacy.

5.3.4 Discussion of factors relating to views on/experience of validity and reliability in EAP testing and assessment

5.3.4.1 Analysis of 'Respondents' views on aspects of construct validity'

This grouping of items, from AQ16 D and E and AQ19 E and F, collectively represent aspects of construct validity (Messick, 1989) in terms of the extent to which respondents agree that students studying different subjects and academic fields should have EAP assessments created which reflect the constructs and idiosyncrasies particular to their subject area.

41.2% of respondents either agreed or strongly agreed in response to items in this series, inferring that they were in favour of creating assessments which represent particular constructs common to specific domains of academic study. In contrast, nearly 30% of respondents either disagreed or strongly disagreed with the need to enhance construct validity in this manner. This left 31% who neither agreed nor disagreed. This distribution of opinion could be said to reflect the contrasting views as expressed by EAP practitioners such as Spack (1988) and Dudley-Evans and St John (1998) and which has been referred to more recently in Hyland (2006). This opinion echoes ongoing professional disagreements surrounding what should constitute the domain for EAP and whether there should be an emphasis on ESAP or EGAP.

As the results of the questionnaire seem to show a mixture of opinion on this aspect of construct validity, similar to that encountered in the EAP sector, perhaps the best recommendation to emerge from this area of the analysis could be in the form of warnings issued by language testing researchers interested in EAP assessment (Fulcher, 1999; Widdowson, 1983). These experts have criticised the traditional focus and debate in EAP surrounding specificity in terms of test content as they believe it has sometimes caused distraction from other aspects of construct validity.

5.3.4.2 Analysis of 'Respondents' views on methods of enhancing validity in EAP testing and assessment'

Although this factor is the product of only two combined items, the researcher's professional judgement along with the Factor Analysis and reliability estimates have led to the inclusion of this factor. The two combined questions (A and E), drawn from AQ20, relate to means of enhancing construct validity of EAP tests through research into EAP students' areas of academic study and collaboration with other colleagues to determine which constructs should be tested.

The findings showed that 87.3% of teachers were in favour of the activities mentioned, thus suggesting confirmation of their agreement for these methods of enhancing the validity of inferences drawn from EAP test scores.

The results of this question can be considered in conjunction with responses analysed in Section 5.3.4, on the topic of 'Respondents' views on aspects of construct validity'. It can be seen that although there is not a majority support for the creation of EAP tests which reflect students' specific academic domains (Weir, 1983), the majority of teachers *do* see the benefits in

researching the features of different subject areas which students are studying and working more closely with academic colleagues from different subject fields (Fulcher & Davidson, 2007, p.86).

This represents an aspect of one of the key steps towards Assessment Literacy as supported by Stiggins (1995) who describes the importance of having full knowledge of what is being assessed. Such behaviour is also linked to more advanced features of teacher ability such as those described by Wallace (1991) as the reflective model, combining classroom practice experience with research. In terms of EAP Assessment Literacy, research and collaboration with academic colleagues in different fields can be viewed as part of the process of identifying a clear criterion which can then be harnessed as a means of reference for both test developers and other stakeholders of EAP assessments.

In the researcher's view it is also positive to note that a large proportion of the respondents recognise the value in interacting with academics in the fields which their students are working towards. This contradicts with the concerns described by some EAP teachers as 'the butler's stance' (Raimes, 1991), inferring a sense of subordination through similar academic interactions with colleagues. It is the researcher's view that, in the best interests of Assessment Literacy, EAP teaching and assessment should be closely connected to the subject areas onto which EAP students progress, given the important gate-keeping role which assessment within EAP courses undertakes (Flowerdew & Peacock, 2001, p.192).

5.3.5 Discussion of factors relating to practices involving analysis and interpretation of the results and scores of EAP tests and assessments

5.3.5.1 Analysis of 'Respondents' views on Criterion/Outcomes based evaluation and feedback in EAP'

Once again the scope of the factor identified from AQ22 is limited as it is based on only two combined items (D and E) from the Likert scale. Nevertheless, the factor has been considered relevant to the study and has been retained after the researcher's consideration of Factor Analysis, reliability statistics and professional judgement relating to the thematic pairing. Item AQ22A collects respondents' comments on connecting student assessment with specific learning outcomes. The second question, AQ22E, relates to the provision of detailed feedback to other stakeholders, in order that test results can be better interpreted by institutions and those who inspect test scores.

The evidence gathered from the questionnaire shows that 86.7% support an outcomes-based approach to assessment (Driscoll and Wood, 2007, p.4) and feedback in EAP assessment whilst only 11.4% neither agree nor disagree and just 1.9% disagree. In a similar manner to the factor discussed in section 5.3.4.2, this represents a positive indicator for Assessment Literacy amongst respondents. A positive attitude towards the outcomes-based approach, such as that demonstrated by responses to the items in this section, can be considered to demonstrate a high frequency of concern for familiarity with assessing the achievement in focus, one of the key aspects of Assessment literacy as identified by Stiggins (1995).

5.3.5.2 Analysis of 'Respondents' familiarity with descriptive and inferential statistics in EAP assessment'

Both the factors stemming from AQ23 will be analysed in this section of the discussion given that the factors represent two sides of the same coin, namely the use of descriptive and inferential statistics.

Factor 1, mainly relating to items from AQ23, involves respondents' views on their familiarity and proactive use of descriptive statistics. The findings demonstrate that the majority of respondents, 67%, are familiar with the use of descriptive statistics however that only 32.9% of those are very familiar insofar as they have used descriptive statistics in their own professional practice. Notably, 20.5% were either unfamiliar with the use of descriptive statistics or very unfamiliar. The remaining 12 % of respondents declared that they were neither familiar nor unfamiliar.

It is perhaps unsurprising that only just under a third of respondents reported high familiarity and use of descriptive statistics, given that, in many situations, the professional skills developed by EAP teachers contrast with those of statisticians. Language teaching professionals may even have chosen their particular career path to side-step more scientific disciplines (Bachman, 2004, p.ix). Despite this situation, the number of respondents who have not been able to report experience in using descriptive statistics in the analysis of test scores, does identify a concern in the researcher's mind. There is a need for more scientific forms of analysis (Bachman, 2004) to complement professional judgement and intuition (Gorard, 2001, p.1-6) in the interpretation of test scores and underlying patterns.

With regard to inferential statistics, notably 62.1% of respondents claimed to be either unfamiliar or very unfamiliar with such procedures, leaving only 16.1 % with some level of familiarity and

9.3% having actually used inferential statistics during the course of their own assessment practice. This outcome is perhaps unsurprising given what the data demonstrates regarding descriptive statistics, which are perhaps more likely to be familiar to lay statisticians through general numeracy and mathematical studies (Healey, 2012, p.7).

Nonetheless, an analysis of the data might imply that that the majority of respondents have a widespread unfamiliarity with the use of inferential statistics, this also suggests that an important opportunity is being missed for the provision of quantitative evidence in order to support test use (Bachman, 2004, p.x). Data of this type can provide deeper investigation than descriptive analysis (Dörnyei, 2003, p.115; Larson-Hall, 2010, p.45).

In the light of the high stakes surrounding EAP assessment, this aspect of Assessment Literacy is particularly valuable as it can support practitioners in demonstrating whether or not tests or assessments are reliable and therefore can be used to diminish risks associated with poor decision making on the basis of test scores and the impact which this can have on individuals.

As recommended by McNamara & Roever (2006, p.8) and Shohamy (2001, p.102) it is important for assessors to think about the consequential applications of their tests beyond the point of operationalisation, in addition to the more routine focus on stages of test development.

5.3.6 Discussion of factors relating to consideration of ethics in EAP testing and assessment

5.3.6.1 'Respondents' views on high-stakes EAP assessment dilemmas 'and 'Ethics in Practice in EAP Assessment'

The remaining two factors based on composite Likert scales relate to ethical considerations associated with EAP assessment and testing. In many ways, principled EAP assessment and the pursuit of Assessment Literacy can be seen to overlap most noticeably in their quest for ethical outcomes. Fulcher and Davidson (2007, p.xix) make the point that assessment should be more than a technical activity but rather an ethical enterprise. In addition to the clear overarching ethical aims of champions of Assessment Literacy such as Stiggins (1995), specific reference is made to the negative consequences of unsound testing practice (Popham, 2001, p.15).

The findings emerging from the first factor, drawn from AQ20 A, B, E and F, which considers respondents' views on high-stakes EAP, indicates that 62.4% of respondents agree or strongly disagree in response to the dilemmas. Analysis of this data could be taken to imply a stance which supports the needs and best interests of students or test takers.

Only 12.1% of respondents recorded their disagreement and arguably adopt a stance which can be argued not to be in the best interest of test takers. 25.4% of respondents neither agreed nor disagreed.

It should be noted that whilst the researcher designed these items with the view that responses would show practitioners' understanding of the need to support the test-takers' best interests, the scenarios which have been included are complex and involve decision-making which may

contradict the advice or position taken by employers. Consequently, this may explain why 25.4% of respondents found it difficult to choose an answer without further contextualised detail.

The researcher considers that the results can be interpreted in an encouraging manner as they can be seen to show that the majority of respondents favour practices which prioritise positive outcomes for test-takers. Such a stance echoes the principles of researchers such as Shohamy (2001) who probes the socio-political implications of language testing. Similarly, Bachman (1990, p.279) acknowledges this view when he refers to the fact that 'tests are not developed and used in a value-free psychometric test-tube; they are virtually always intended to serve the needs of an educational system or of society at large.'

In terms of the construct for Assessment Literacy used to steer this investigation surrounding EAP, the concern for ethics can be considered as a fundamental aspect of building awareness of what can go wrong in EAP testing and how problems can be prevented before they arise (Stiggins, 1995).

A key aspect of Popham's mission in the promotion of Assessment Literacy (2001, 2006, 2009, and 2012) appears to be his assertion of the need for proactivity in order to generate change, through a series of action challenges advocating practitioner involvement in the process of enhancing Assessment Literacy. With this action-based approach in mind, the next factor, taken from AQ26 E, F, G and H, asked respondents to comment on areas of ethicality in their own practice and employment.

41.9% of respondents returned an answer which, when analysed, implies that they feel that their work and employment in association with EAP has not been connected to unethical practices. 20.9% of respondents responded in a manner indicating that that they do feel that their work and

employment has been connected with unethical practice. This left 37.2% of respondents who selected the 'unsure' option.

When compared to the previous factor, this finding raises more concerns and suggests that whilst Respondents appear to have the best ethical interests of test-takers at heart, only just over 41% of respondents felt able to declare that their employment and practice was not associated with unethical practice.

The researcher's interpretation and recommendation on the basis of these findings would be the provision of additional information or training to the sector so that those who are unsure about what might constitute unethical practice can be better informed. Similarly, those who feel certain that that they have been involved or associated with unethical practice have a mechanism to address the issue or gain support for their concern outside the boundaries of any particular institution. If this is not the case then dangers for stakeholders may ensue. Warnings about the negative implications of language testing are provided by Spolsky (1981, p.20) who states that language tests test should be given a health warning similar to those given to dangerous drugs or chemicals. Furthermore, test ethicality is such a priority for some educationalists (Lynch, 1997; Lynch & Shaw, 2005) that this has led to alternative approaches to language testing in order to avoid the power imbalances which some practitioners believe are characteristic of traditional testing systems.

5.3.7 Section conclusion

This area of the analysis has identified some of the most significant features of the data which has been collected, through responses to clusters of related questions in certain key areas of assessment literacy. The implications of these findings have been disussed in relation to what they

reveal about the Assessment Literacy of EAP teachers. An additional level of detail can be added to what has been determined in this section through analysis and discussion of responses to indivudal items, which follows in Section 5.4.

5.4 Discussion and analysis of Set A2 - Descriptive statistics for individual questionnaire items

5.4.1 Section introduction

Individual findings have been presented in the form of descriptive statistics for each of the DVAS items which were used in the different sections of the questionnaire and composite results have been provided, in section 5.3, for groups of related items have been analysed. As a result, only those individual items which have not been included in the composite results, but which the researcher considers add additional and important information, will be discussed in this section.

5.4.2 Discussion of results linked to training, skills, strengths and weaknesses in EAP testing and assessment

5.4.2.1 Analysis of individual DVAS items from AQ13

The three questions which will be considered in this section stem from AQ13 and discuss respondents' interest in EAP, perceptions regarding the fundamental importance of assessment skills in EAP practice and the means through which respondent skills have been developed.

Item AQ13a reveals information regarding respondents' level of specific interest in EAP assessment. Notably, 87.3% or respondents (138 individuals) either agree or strongly agree that their interest in EAP is driven by aspects of the field other than testing or assessment. It is perhaps

unsurprising that many educators may embark on a teaching career for reasons other than an interest in assessment. Nevertheless, it should be noted that the promotion of practices associated with Assessment Literacy require a level of dynamism in order for action to take place which will result in better tests and better judgements based on tests. The researcher argues that in order for Assessment Literacy to benefit from movements such as assessment for learning, 'the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there.' (Assessment Reform Group, 2002), the interest and attention of a sufficiently large number of practitioners is required.

Popham (2009, p.11) acknowledges the contribution of assessment for learning to Assessment Literacy by stating that, when classroom assessments are conceived as assessments for learning, rather than assessments of learning, students will learn better what their teacher wants them to learn. However if EAP teachers and assessors are not sufficiently interested in matters associated with assessment, there is a danger that the benefits to assessment may not be maximised.

When considered in conjunction with the results of AQ13b, which shows that 80% of respondents recognise the fundamental importance of assessment in EAP, it would seem advisable to consider mechanisms for assisting EAP tutors in deepening their level of interest in assessment, so that the benefits of further attention to this aspect of pedagogy can be reaped.

Interestingly, when the results of AQ13c are reviewed it shows that only 30% of respondents agree that their EAP assessment related skills have been developed through courses with a focus on language testing. This finding is corroborated by Sharpling (2002), who suggests that the problem with courses used for EAP training purposes is that EAP teachers are required to acquire knowledge which is not directly available via training courses. Although Brown and Bailey (2008)

recognise that the matter of training for classroom testing and assessment has long been considered important in the context of general education, a deficit in teacher assessment skills is highlighted by Stiggins (2002, p.762) who explained that, 'Few teachers are prepared to face the challenges of classroom assessment because they have not been given the opportunity to learn to do so.'

5.4.2.2 Analysis of individual DVAS items from AQ14

Whilst the composite Likert scale results make use of all items from AQ14, which represent respondent skills in specified areas of EAP testing and assessment, it is useful to consider patterns displayed across the individual items given that in additional to overall skill markers, individual aspects of sub skills within EAP assessment are relevant to the study.

When considering the findings it appears that that 9 of the 13 items show a clear modal response in the 'Average Skilled' category suggesting that the development of skills across a wide range of capabilities associated with EAP assessment would be of benefit to practitioners and the sector.

When evaluating the skill areas in which respondents appear to have most confidence, the skill areas in which respondents rate themselves highly relate to giving feedback to students and identifying what needs to be assessed, showing that respondents feel aware of student needs and feel skilled in advising students on their learning. These represent key factors in both Assessment for learning and Assessment Literacy. Other elements in which fewer respondents were able to claim high skill, involved use of statistics, defining test purpose, and trialling and piloting tests. A total of 7 areas showed less than 40% of respondents were able to rate themselves as highly skilled thus indicating key areas for consideration in training or development opportunities.

5.4.3 Discussion of results linked to approaches to EAP testing and assessment design

5.4.3.1 Analysis of individual DVAS items from AQ16

AQ16B is focused on the use of test specifications and although this item did not cluster usefully with other items it is regarded as pertinent by the researcher. It was noted from the findings that 64.5 % (102) of respondents reported agreement or strong agreement with the need to use a test specification in the process of test design. Whilst this represents the results of just a single item, it can be taken to imply a level of awareness of test specifications as important guidance tools in the process of EAP test development, which can be considered a feature of assessment. According to Davidson and Lynch (2002, p.3) the test specification is the primary tool for language test development.

When evaluating respondents' views on collaboration with academic content tutors in the process of EAP test design, AQ16c demonstrated that 72.1% (114) of respondents disagreed or strongly disagreed with the view that the involvement of academic content tutors was unhelpful. Whilst Factor Analysis and reliability statistics did not show a grouping of this item with the two items which were combined in the composites from AQ20 as shown in section 5.3.4.2, both the theme and magnitude of responses compare favourably, again showing that the EAP teachers who completed the survey value working with academics from other subject areas.

Item AQ16h shows respondents views regarding different testing and assessment situations and whether there is a need for different marking descriptors in each context. Responses showed that 65.1% (103) of respondents agreed or strongly agreed with the need to use different marking descriptors for different testing situations.

A level of comparison can be drawn with findings from this item and the set of items in 5.3.4 which relate to aspects of construct validity connected to the representation of different domains of study. Although, through the evidence in section 5.3.4 only 41.2% of respondents either agreed or strongly agreed with the need to create assessments which represent particular constructs common to specific domains of academic study, 65% of respondents to AQ16h report a need for different marking descriptors in different contexts. This difference could partly be explained by respondents' views which may infer that assessment contexts differ within subjects as well as across subject areas leading to a starker sense of need to consider particular assessment tasks rather than the need to consider subject specialism.

AQ16h results can also be linked to the results in 5.3.5 which refer to views on outcomes-based assessment (Bloom, 1956; Anderson & Krathwohl, 2001). Both AQ16h and the response items drawn from AQ22 show a majority level of support for an outcomes-based approach to assessment and feedback in EAP assessment.

The analysis of other items from AQ16 has not been presented as they have been considered ambiguous or insufficiently meaningful.

5.4.3.2 Analysis of ranking items from AQ17

When considering the results of AQ17, which required respondents to rank key stages of EAP test and assessment design and use, the overall results demonstrated a pattern which was consistent with a good level of understanding amongst respondents of logical and principled staging and sequencing in EAP test design and usage.

Although the series of stages was based on the Test Design Framework (Fulcher, 2010, p.94), a number of amendments were made including the addition of item AQ17H, as a potentially less critical stage in test design and the additional reference to the use of statistical procedures.

As item AQ17H was the most frequently omitted item, this showed that a significant number of respondents identified it as a less critical element of the test design process. Items AQ17G, D and E, were the series of items with the next highest rate of omission. These items relate to the identification of main skills, sub skills and the creation or consultation of a test specification.

It can therefore be suggested that it may be worthwhile providing additional guidance for teachers of EAP on the benefits of giving attention to both macro and micro language skills in the process of construct operationalisation as well as ensuring that practitioners are aware of the useful guiding function of test specifications in maintaining construct validity.

5.4.4 Discussion of results linked to views on/experience of validity and reliability in EAP testing and assessment

5.4.4.1 Analysis of individual DVAS items from AQ19

Whilst the items in this section are intended to focus on respondent views on aspects of construct validity, the series of questions in AQ19 have been shown through validity and reliability analyses not to congregate except in the case of two items. On rereading the list of questions, the researcher judges that this has occurred due to a lack of precision in the phrasing of some items. The situation has also occurred due to the identification of a spectrum of issues connected to construct validity which is too broad for commonalities necessarily to occur. Nevertheless, a number of individual items have still been identified as worthwhile for consideration.

The results of AQ19A show that 46.2%, nearly half of respondents agree to some extent that it is difficult to be specific about what constitutes a particular area of linguistic skill in EAP. This finding compares favourably in terms of theme with the results which emerge from AQ13, items H, I and J which refer to views on the complexity of EAP assessment as AQ19 A also suggest that a significant number of respondents find it difficult to identify which aspects of a particular language skill, or construct, to assess in EAP assessment. This also concords with the views of experts in the field of language testing who recognise the challenge posed by operationalizing language constructs in assessments (Bachman & Palmer, 1996, p. 62; Fulcher & Davidson, 2007; p.7).

The results relating to AQ19B are useful in terms of what they reveal about respondents' feelings with regard to large scale standardised testing connected to EAP, such as IELTS and TOEFL (Alexander et al., 2008, p.308). Results demonstrate that the majority of respondents, 62.6 %, disagree or strongly disagree that results from IELTS or TOEFL are more likely to be valid and reliable than those based on the results of their own EAP assessments. Whilst this is a view shared by members of the BALEAP working party on testing (2012), including Schmitt (2012), this view does not appear to be shared by the UKBA given recent policy changes which now require Secure English Language Tests (SELTs) in association with Tier 4 study visas.

Other items which have not been covered in this section are considered too ambiguous in their structure and findings to warrant further coverage.

5.4.4.2 Analysis of individual DVAS items from AQ20

One particular item from AQ20 which was not part of the composite analysis which the researcher believes warrants further discussion is AQ20B. The results of AQ20B showed that 81.6% of

respondents felt that incorporating multiple measures of a construct would assist in enhance the validity of inferences drawn from test scores. Analysis of the data from this response might be taken to imply respondents' understanding of the importance of considering parallel forms of assessment which aim to measure the same construct (Fulcher and Davidson, 2007, p.105). This is an important factor in the process of improving reliability in language testing.

5.4.5 Discussion of results linked to practices involving analysis and interpretation of the results and scores of EAP tests and assessments

5.4.5.1 Analysis of individual DVAS items from AQ22

Discussion related to individual items within AQ22 indicates through consideration of items A and B that, in both cases, approximately 55% of respondents refer to the difficulty of both analysing and drawing concrete conclusions from EAP tests. The expression of this level of concern amongst more than half the respondents may suggest the need for further guidance mechanisms to assist in both test design and the analysis of results which follow.

The results of AQ22G indicate that a significant number of respondents feel a strong sense of responsibility for ensuring that other users interpret EAP test results appropriately, however the results of AQ22F suggest that nearly 70% of respondents are not able to agree that they frequently work with other academics to help explain EAP. This situation suggests a need to action measures which allow or encourage EAP teachers to play a more proactive role in liaising with academic colleagues in this manner so that stakeholders have better understanding of EAP test scores.

The final question which will be considered here is AQ22J which relates to views on the impracticality of converting EAP scores into IELTS or TOEFL equivalents. The results demonstrate that whilst 44.4% of respondents recognise the impracticality of generating IELTS- or TOEFL-style scores from in-house testing systems the remaining 55.6% of respondents are either unsure or in agreement with the creation of such equivalences. The researcher considers these findings alarming given that both the test specifications and grading criteria for the standardised tests mentioned are not widely available in the public domain and that emulation of the idiosyncratic grading systems risks inferring a level of parity which cannot be proven. Consequently this raises concerns relating to criterion referencing and the use of grading criteria in a principled and transparent outcomes-based approach.

5.4.6 Discussion of results linked to the consideration of ethics in EAP testing and assessment

5.4.6.1 Analysis of individual DVAS items from AQ25 and AQ26

The researcher does not feel further attention is warranted regarding individual items in AQ25 and AQ26, which relate to additional ethical scenarios surrounding EAP. This partly due to the fact that the topic of ethics has been covered successfully through the items which clustered reliably in AQ25 and AQ26 and also due to the fact that a number of the items which did not cluster are now not considered sufficiently meaningful, given aspects of ambiguity which are inherent to the questions themselves.

5.4.1 Discussion of results linked to influence of research and other resources for the purpose of EAP testing and assessment on EAP teacher assessment practices

5.4.1.1 Analysis of individual DVAS item AQ13d

As mentioned in the Findings section, although AQ13d is single item, which reveals information about respondents' reference to research undertaken on the topic of EAP testing and assessment, it can be considered relevant when combined with relevant results from the interview. Only just over one third (34.8%) of respondents agreed or strongly agreed that they consulted research-led resources associated with EAP assessment with regularity. This does suggest that intervention is necessary which can assist practitioners in accessing and learning from research more proactively.

5.4.2 Section conclusion

Analysis of the results of individual questions, which have not been considered during the analysis of clusters of items, has enabled the identification of additional pertinent information relating to the exploratiuon of EAP teacher Assessment Literacy. This information can be considered further and synthesised alongside the details which emerge from other sets of data, in order that the full set of data collected can be considered to have been explored as fully as possible and that frequency of response to individual items are not unnecessarily disregarded.

5.5 Discussion and analysis of Set A4 - Open questionnaire responses

5.5.1 Section introduction

As the open responses to the questionnaire were amongst the few questions in the online survey tool which were marked as optional, the number of responses returned does not reflect the magnitude of response to the other sections of the questionnaire. Nevertheless, whilst clusters of commonality are comprised of fewer individual cases, the implications of patterns observed will

still be discussed as they may be relevant in terms of support for observations which emerge from other areas of the questionnaire and the interviews.

5.5.2 Analysis of open question AQ15, AQ18, AQ21, AQ24 and AQ27

After review of the main factors which emerged from AQ15, the focus of which was on EAP training and skills development, it can be noted that the clearest pattern of shared opinion relates to a lack of specific or specialist training in EAP assessment. 11 respondents added comments which infer this experience. This finding also connects with the comments made by Sharpling (2002) who identifies the fact that EAP teachers need to acquire knowledge which is not directly available via training and that whilst the series of qualifications available to EAP teachers maybe broad, it is difficult to find parity across the range, in particular with regard to practical components. This finding also links with findings gathered through AQ13c which shows that only 30% of respondents agree that their EAP assessment related skills have been developed through courses with a focus on language testing.

The most striking observation to emerge from AQ18, on the topic of EAP assessment and testing design, can be gleaned from a grouping of 11 responses which indicate that respondents consider the type of process adopted for test design to be dependent upon the particular purpose and setting for which the test is being created. In terms of Assessment literacy, this view resonates with the view held by Stiggins (1995), regarding the importance of assessors having knowledge about the purpose of testing and assessment. In relation to EAP this is corroborated by Alexander et al. (2008), who summarises the varied purposes of tests whilst connecting them to EAP context.

With respect to AQ21, rather than providing a particularly meaningful view of shared opinion on the topic of construct validity, the responses which cluster most noticeably for this item highlight

flaws in the questions themselves. The most common cluster of responses given shows respondents understanding that validity in EAP testing is highly dependent on a wider range of factors than those presented in the questions in the questionnaire.

The number of thematic clusters and rate of response to AQ24, on the topic of interpreting and analysing EAP test results, is notably lower than open items which occurred towards the start of the questionnaire. The factor with the highest number of common comments is also remarkably low. Factor 1 from AQ24, with just 4 common responses, refers to difficulties experienced by respondents associated with IELTS, in terms of the inefficiencies of interpretation of scores for admissions, and the dangers of creating IELTS-equivalent scores for in-house tests.

Finally, the main pattern of response to AQ27 is manifested in the form of ethically oriented statements relating to respondent experience or moral positions associated with EAP testing and assessment. The 8 comments provided show a keen level of concern for practice which prioritises the needs of stakeholders rather than institutional requirements (Benesch, 2001; Shohamy, 2001; McNamara & Roever, 2006).

5.5.3 Section conclusion

Although it is recognised that the volume of responses to the open questions which were analysed in this section did not match that of other items which were marked as compulsory, it is clear that a number of noteworthy patterns of response have been identified. It has also been useful to analyse this data so that the implications can complement other analyses which have been undertaken. It could also be argued that despite the lower volume of clustered responses, discernable patterns are nonetheless valuable, given the spontaneity of their provision by questionnaire respondents to a relatively unguided open question.

5.6 Discussion and analysis of Set B3 - Interview content analysis and interactional profiles

5.6.1 Section introduction

In this section, the major patterns of meaning which emerged from the interview stage of the research will be analyzed and discussed in relation to the research questions for this study, surrounding the Assessment Literacy of EAP teachers. In addition, interactional profiles which relate to each interview will be considered, with a view to examining the nature of interactions between the interviewee and interviewer. It will then be considered whether this reveals any additional information which needs to be taken into account when the interpretation of the data from the interviews is taken into account.

5.6.2 Discussion of results from interview content analysis

5.6.2.1 Analysis of interview factors linked to training, skills, strengths and weaknesses in EAP testing and assessment

In this first pattern of response, questions form BQ2-BQ5, relating to skill development in EAP testing and assessment will be considered alongside interviewees' views on their strengths, weaknesses and greatest challenges.

From BQ2 the most notable pattern connected to the development of skills in EAP testing and assessment was expressed with regard to the development of EAP-related skills predominantly on-the-job (Sharpling, 2002), rather than through specific training, with 20 interviewees making this observation. Whilst a second prevalent factor did reveal that a number of respondents were able to

acknowledge forms of specific related training, this was only 9 interviewees. This pattern of response correlates with findings from other aspects of research, namely Question AQ15 and AQ13C which both reveal that teachers feel they have acquired their skills in EAP assessment through their work rather than through their education.

With respect to responses associated with strengths in EAP assessment (BQ3), factors which emerged from the data showed respondent confidence in: gauging student learning needs, the importance of their own academic background in terms of subject knowledge or ESAP and the act of producing EAP test materials. The information emanating from the 25 interviews showed a wide variety of perceived strengths, with very little reference to the analytical or interpretive stages of testing and a more evident focus on test design. Given the need to create EAP tests which are used for gate-keeping purposes (Flowerdew & Peacock, 2001, p.192) and the individuals whose futures may be at risk, if practitioners are not able to refer to strengths in test score interpretation, this may suggest the need for further attention in this area.

Given that the main factor to emerge from BQ4 expresses interviewees' perception of a general lack of competence or experience in EAP testing and assessment, the requirement for further training needs is again highlighted. Similarly, suggestion that additional training may be warranted is further supported by the results of BQ5 relating to interviwees' biggest challenge. The factors which arose from this particular question and which showed the most prominence were connected with concern to undertake assessment well, or 'to get it right' and to assess EAP in a manner which reflects students' real needs in the University context. This result suggests the need to support practitioners with models of good practice and training.

5.6.2.2 Analysis of interview factors linked to approaches to EAP testing and assessment design

The results of questions connected to principled processes followed in EAP test design and the means by which interviewees identify or create suitable test items was the focus for BQ6 and 7.

As acknowledged in the results section, whilst a number of smaller groupings of commonality between interviewees do indicate set approaches to EAP assessment and test design, the most striking observation can be related to BQ6 where 12 interviewee responses mentioned the absence of an identifiable procedure in their practice of EAP test and assessment design. In terms of Assessment literacy, this suggests the need for guidance on principled test design.

With regard to the identification of suitable questions for use in EAP tests, the factor which emerges most prominently from BQ7 links to comments associated with the need to identify test questions which link to intended learning outcomes. This shared view amongst interviewees could be considered to imply a strong understanding of the benefits of criterion-referenced assessment which can be linked to Stiggins' (1995) view of the importance of having full knowledge regarding what is being assessed as well as Bloom (1956) and recommendations for good practice in EAP Alexander et al.(2008). This can also be connected to the analysis of factors emerging from AQ22 in section 5.3.5 which similarly show that respondents' questionnaire responses support an outcomes based approach.

5.6.2.3 Analysis of interview factors linked to views on/ experience of validity and reliability in EAP testing and

The analysis of responses to items BQ8-10 relates to questions associated with validity and reliability of test scores and domain representation in EAP testing. Notably, with regard to BQ8 and BQ9, the factor with the clearest grouping of shared response is characterised by interviewee

lack of confidence with validity and reliability in EAP testing and assessment. Again this raises questions and concerns regarding the level of training available to EAP practitioners (Krzanowski, 2001; Roberts, 2001; Sharpling, 2002) involved with assessment and testing. An intervention seems to be required to enhance the understanding of validity and reliability in EAP testing.

Other factors in which groupings of shared concern are expressed by interviewees, are associated with the need to undertake trialling and piloting, in addition to the importance of moderation in order to improve the validity and reliability of inferences drawn from test scores. In terms of Assessment Literacy, this understanding, expressed by interviewees recognises that principled language testing in the EAP context should be an iterative and organic collaborative process (Fulcher and Davidson, 2007, p.61).

As far as opinion relating to domain representation in EAP assessment is concerned, Factors 1 and 4 from BQ10 emerged with the highest incidence of related responses from interviewees. The results of these factors suggest that interviewees attach importance to choosing subject matter in tests connected to the academic field of students' study (Weir, 1983) as well as working in collaboration with academics from the relevant academic fields.

Comments relating to working in collaboration with other colleagues can be considered relevant in a similar manner to the results in the composite results discussed in section 4.4.11 concerning processes and procedures in EAP assessment. The composite results of AQ16, relating to processes and procedures in EAP testing, also show a strong level of support for the positive effects of teamwork (Davidson & Lynch, 2002; Jafarpur, 2003, p.72). A similar overlap is observed with the findings of the pairing of items from AQ20 A and E which relate to methods of

enhancing validity in EAP testing and assessment and where the importance of collaboration with other academic colleagues is also acknowledged.

One concern which does however also emerge from the findings of BQ10 is the reference by a number of interviewees to barriers associated with working with colleagues in other departments across institutions. As a result means could be sought to facilitate, encourage and promote the activity as described above. This represents a key aspect of assessment literacy in terms of collaboration to understand the filed of EAP more fully and to enhance validity.

5.6.2.4 Analysis of interview factors linked to practices involving analysis and interpretation of the results and scores of EAP tests and assessments

With respect to interpretations and analysis of results from EAP test scores, BQ11 shows that nearly half of interviewees referred to the need to use grading criteria. This presents another example of reference to good practice associated with outcomes-related assessment (Anderson & Krathwohl, 2001; Driscoll & Wood, 2007, p.4). However it is notable that 5 interviewees made comments which could be taken to imply the unscientific nature of their interpretations and although one or two individuals did comment on their use of statistics, this was not a prevalent theme amongst those interviewed.

Analysis of BQ12 shows that one of the more noticeable groupings of opinions, with 6 similar comments from interviewees, again surrounds the less-than-scientific mechanisms practiced by EAP teachers and assessors in the process of interpreting scores, thus inferring the need for training in this area to build EAP Assessment Literacy. Seven comments related to the use of moderation in maintaining accuracy in the interpretation of EAP tests and assessments. This

agains suggests an understanding of the powers of collaborative working and consensus (Davidson & Lynch, 2002; Fulcher and Davidson, 2007, p.61). This is critical in EAP Assessment Literacy, in order to understand what is being tested as fully as possible.

The results of BQ13 reveal EAP teachers' experience of assisting stakeholders in the understanding of test scores. This is part of addressing the needs of EAP test stakeholders' an an element in the process of developing of EAP Assessment Literacy. Whilst the most prominent grouping of comments relates to the provision of detailed formative feedback to students, a cluster of comments is also observed regarding the difficulties associated with IELTS and comparing grades from in-house tests to IELTS bands. This latter comment resonates with the results from AQ22J where the dangers of creating such roughshod equivalences are described despite a split in practitioner views on their practicality.

5.6.2.5 Analysis of interview factors linked to the influence of research and other resources for the purpose of EAP testing and assessment

The main factors emerging from this section of the interview, taken from responses to BQ14-16, focused on the influence of published research on EAP teachers and their assessment practices along with the need for any other resources. Analysis of responses to BQ14 shows a clear split of interviewee opinion with 8 interviewees suggesting explicitly that they were markedly influenced by research and 12 stating that they were not. Whilst it is positive that some interviewees noted an influence, the number of interviewees who recognised a lack of influence of research is alarming given the distance which this places between practitioners and the key principles which have been published in the growing literature surrounding language testing which is relevant to EAP. This suggests that there is a distance between EAP teachers and the expanded Assessment Literacy

which they require (Taylor, 2009). This situation also represents a training opportunity which the researcher considers to be crucial.

A divide in opinion is noted from responses to BQ15 which shows that whilst 5 interviewees refer to specific titles of books or journals which they refer to and 4 speak more generally of material they have consulted, the most marked convergence of opinion shows that 11 interviewees do not refer any particular books or journals. Once again, the researcher considers this situation to require further attention given that it has the potential to restrict the implementation of each of the key areas of Assessment Literacy as discussed in the literature review.

Analysis of BQ16, which gathered information on the topic of additional resources requested by interviewees for assistance in EAP assessment and testing, revealed a grouping of responses from five interviews indicating a need for additional assessment-related training. Although this finding is based on a small area of convergence in the group of interviewees, it supports other findings from this research project, which can be considered to imply the need for additional training.

5.6.2.6 Analysis of interview factors linked to consideration of ethics in EAP testing and assessment

When considering the most prominent results stemming from BQ17, it is apparent that interviewees most frequently referred to concerns connected to reliability in EAP testing and more specifically, perceptions relating to the unreliability of large-scale standardised testing in the form of IELTS. In terms of Assessment literacy, these issues highlight interviewees' understanding of the dangers to both institutions and individuals if unreliable measures are interpreted and issued. This analysis supports the findings drawn from the questionnaire which also suggest that respondents have demonstrated an ethical stance which aligns with good practice recommended

by supporters of Assessment Literacy and critical language testing (McNamara and Roever, 2006; Popham, 2001; Shohamy, 2001; Stiggins, 1995).

Whist BQ18 was also designed with a view to collecting responses associated with ethical concerns, the phrasing of the question was clearly too ambiguous. This led interviewees to make reference to idiosyncratic features of EAP teaching, learning and assessment, rather than considerations relating to ethical implications of EAP testing and assessment. Consequently, the largest cluster of opinion expressed referred to the different genre of English which EAP represents and the need to assess a different set of skills associated with spheres of academic practice or particular academic domains. Whilst this type of response was not that expected by the researcher, it does identify a trend and a level of practitioner understanding which recognises both student and institutional needs. Reference to domain also resonates with the early research of Weir (1983) who contended that content validity improved in the assessment of EAP through analysing test takers needs in relation to their courses of study and using this information to influence the content of tests.

5.6.1 Analysis of interactional profiles

Through examination of the 25 interactional profiles completed in relation to the interviews, it has been noted that interview duration ranged from a 5 to 47 minutes with the average interview duration being 17 minutes. As Table 130 below indicates, the majority of interviews had a duration of between 11 and 15 minutes.

Table 130 Interview duration in whole minutes

Number of Minutes	Number of interviewees
5-10	4
11-15	10
16-20	5
20+	6

With regard to the relative status of the interviewer and interviewee, it should be noted that the interviewer made concerted efforts to ensure that all interviewees felt comfortable and that an environment was created, whether that be face-to-face or through Skype, where a peer-to-peer interaction could take place. Whilst the interviewer is confident that no situation was discerned whereby interviewee responses were affected due to a perception of the researcher's seniority, there were certain situations where it was noted that the interaction in the interviews recognised the seniority of the interviewee. This situation either arose naturally through the manifestation of a particular interviewee's level of experience in comparison to that of the interviewer or through a sense that the interviewee wished to have their level of seniority recognised by the interviewer.

Through considering the researcher's observations relating to participant status, it can be noted that in 21 of the 25 cases, the researcher felt that a peer-to-peer interaction was maintained. In interviews IG, IL and IY it was identified that the interviewee positioned themselves as the senior partner. In interview IH the interviewee positioned themselves as a junior partner due to the interviewee's apparent sense of inexperience. It should also be recognised that interviewee ID is a junior colleague of the researcher at his current place of work, however, despite this situation, the researcher is confident that a peer-to-peer interaction was maintained at all times during the interview.

As the researcher is confident that each interactional situation allowed interviewees to speak freely without any notable debilitating or restricting effect caused by perceptions of status, no further action has been taken on the basis of this information.

As far as interviewee engagement is concerned, each interviewee's contribution was rated by the researcher on the basis of the researcher's perception of their level of enthusiasm and interest for

the topic during the interaction. Table 131 describes the researcher's interpretation of the group of 25 interviewees:

Table 131 Interviewee level of engagement

Level of engagement with subject	Number of interviewees
Very engaged	5
Engaged	11
Neither engaged nor disengaged	8
Disengaged	1
Very disengaged	0

Five of the interviewees appeared engaged with the topic and seemed to speak with a very high level of interest and confidence on the topic of EAP assessment. Almost half of the interviewees (11) were engaged with the topic and spoke freely and easily about the subject of EAP assessment. Eight interviewees seemed only moderately interested and motivated to talk about the topic, leaving just one interviewee who seemed disengaged on the topic. In considering the nature of interactions, it was also notable that longer interactions did not necessarily yield richer information, as longer stretches often included irrelevant or repeated material.

Concerning other information provided in the profiles, it was noted that a number of interviews were time-limited due to the need for interviewees to attend a meeting or another appointment at a particular time. As a result, it is possible in some cases that this could have restricted responses.

5.6.2 Section conclusion

With the addition of interpretations emerging from the analysis of clustered Likert Scales items from the questionnaire, the analysis of factors arising from the interviews, as presented in this section, demonstrates that this is particularly important information set within this research project. The understandings which stem from this qualitative analysis make a key contribution to what is a 'thicker description' (Denscombe, 2007, p.63) of the Assessment Literacy of EAP teachers and provide a key insight of particular value when answering the research questions in

this study. The relevance of the data collected and analysed from the interviews has been further supported through analysis of the interactional profiles, which indicate that the individuals interviewed represent a suitably relevant and reliable sample of the target population and that data collection from the interview has not been unduly affected through interaction with the researcher.

5.7 Chapter conclusion

This chapter of the thesis has enabled the analysis of key information which has been extrapolated from the findings of research undertaken through the main collection tools used for this study into EAP teacher Assessmet Literacy. Through further evaluation of the data which has been gathered, it has been possible to start to consider the meaning and implications revealed within the findings, in terms of what they indicate about EAP teacher views on their assessment practices and the extent to which this reflects language testing research and practices which comprise Assessment Literacy. Now that an initial analysis of the results of the questionnaire and the interviews have been considered, it is possible to move on to the next chapter where the components of the analysis will be synthesized in the conclusion. Recommendations will also be made for sustaining or enhancing Assessment Literacy amongst EAP teachers, which will help to answer the second research question which drives this thesis.

CHAPTER 6 CONCLUSIONS AND RECOMMENDATIONS

6.1 Chapter introduction

As suggested in final section of the analysis chapter, the conclusion to this thesis will now combine the observations which have been drawn from the analysis of research, in order to identify and group together interpretations which characterise the Assessment Literacy of participants in this study, which can in turn be considered generalizable to the wider sector of EAP practitioners. Observations will be related back to the original research questions and hypothesis, as stated in Chapter 1, with reference to the theoretical frameworks which underpin the field under exploration. The limitations of the findings will also be presented. Finally, recommendations will be made which suggest how the observations from this study can most practically and usefully be implemented. These recommendations will also collectively highlight the important contribution which the researcher believes this investigation has provided to the field of EAP Assessment Literacy.

6.2 Synthesis of observations from the research

The synthesis of observations from this research will be organised in line with the research questions which sought to investigate the extent to which EAP teacher views on EAP testing and assessment practices reflect language testing research and practices which comprise Assessment Literacy. The reasearch objective also included the aim to ascertain how EAP Assessment Literacy could be sustained or enhanced. As with other areas of this thesis, the observations will be organised in line with the Key areas of EAP Assessment Literacy which were identified for exploration in this study during the Literature Review.

6.2.1 Observations relating to EAP teacher training, skills, strengths and weaknesses in EAP testing and assessment

With regard to practitioner levels of interest and confidence in EAP assessment, the results of the study appear to indicate considerably less than favourable levels in both areas, with data, collected from both composite Likert scales and individual items within AQ13, suggesting relatively low levels. In addition, data drawn from the interview stage of the research from BQ11 and BQ12 both show clusters of comments from interviewees which refer to less-than-scientific mechanisms for test interpretation.

The Likert scale items from AQ14, which collectively act as an overall skill-evaluation tool with relevance to Assessment Literacy in EAP, show that respondents position themselves predominantly as 'average skilled' in the key aspects of assessment in EAP which were listed.

When considered individually, it can also be seen that the modal response for 9 of the 13 options is also 'Average Skilled'. Although patterns also emerge where respondents report to be highly skilled in identifying students' needs and advising students on their learning, practices associated with Assessment for Learning (Berry, 2008; Knight, 1994), it can be argued that training opportunities for such skills are already widely available through existing language teaching courses (Krzanowski, 2001; Roberts, 2001; Sharpling, 2002). Other skills such as use of statistics, defining test purposes and trialling and piloting tests, which were identified as areas of lower skill, may not be so easily developed through existing general training opportunities. The results of BQ3 can also be taken to imply support for this situation. In response to BQ3, a pattern of interviewee response described strengths in EAP testing and assessment but largely did not refer to analytical or interpretive stages of test development and usage. Consequently, there is a case for the

provision of accessible in-service or pre-service training which makes such skills more easily available to practitioners.

With regard to the manner in which skills in EAP assessment have been developed, a series of sources from within the data sets, including DVAS item AQ13, free text item, AQ15 and clusters of comments from the interview stage, suggest that skills relevant to EAP testing and assessment have predominantly been developed in the workplace due to a lack of related specialist training available. Whilst it may be the case that some of the skills connected to EAP teaching need to be developed through practice rather than tuition (Sharpling, 2002), in order to improve practitioner confidence and enjoyment with principled EAP testing and assessment, it is worthwhile considering more widespread promotion of access to training or CPD mechanisms related to Assessment in EAP.

A cluster of responses to BQ5 show that practitioners are aware of the need to undertake the task of EAP assessment to the best of their ability, given challenges which they face and the stakes in the balance. However, the results of a Likert scale and an additional individual item show that more than 50% of respondents did not actively acknowledge the high level of complexity associated with EAP testing and assessment. This raises concerns for the researcher as this situation may lead some practitioners to underestimating the amount of effort required in order to create trustworthy tests, based on satisfactorily operationalised constructs (Moss, in Fulcher and Davidson, p.200). This is necessary to allow the drawing of inferences which may be considered trustworthy (Messick, 1989). As a result an awareness-raising activity is recommended which draws attention to the multifaceted and complex nature of EAP test design and usage, whilst simultaneously providing scaffolding to support the time-pressured practising EAP teachers towards whom the intervention is aimed.

6.2.1 Observations relating to the influence of research and other resources for the purpose of EAP testing and assessment on EAP teacher Assessment practices

When data linked to practitioners' consultation of published research is considered, the interviews show a split response, with nearly half the interviewees not consulting research actively, this was supported by results from BQ15 which showed 11 responses which mentioned no reference to specific titles connected to EAP assessment. Results from AQ13d also corroborated this. One conclusion which can arguably be drawn from these findings is the need for an intervention which assists in building practitioner skill in areas associated with EAP assessment in order that confidence can be built and key aspects of Assessment Literacy can be fostered. This could then serve as a means of providing EAP teachers which the expanded Assessment Literacy (Taylor, 2009) which they require, as language teaching professionals. It may be advisable to build into this process a mechanism which proactively introduces practitioners to published research which is relevant to EAP assessment in a manner which is perceived to be user-friendly and accessible, so that the knowledge embedded within these resources can be more proactively absorbed and disseminated.

6.2.2 Observations relating to approaches to EAP testing and assessment design

Analysis of the series of results from across the range of data has been taken to imply that a large proportion of EAP teachers engaged in assessment have a good understanding of what constitutes a principled approach for EAP test and assessment design (See Fulcher, 2010, p.94). The Likert scale results for AQ16 items A, F and G indicate respondents' understanding of test design as a collaborative venture requiring time and key stages in advance of item writing (Bachman & Palmer, 1996, p. 62; Fulcher & Davidson, 2007; p.7). The free-text open responses to AQ18 show

that respondents see test design as a highly context-driven process based on the particular purposes for which the test is required. Responses to AQ17, which ranked stages of test design and use in terms of priority, showed a pattern which was consistent with logical and principled sequencing of stages of test design. Nevertheless, despite practitioners' ability to recognise a principled approach, the findings from the interview stage in item B6 showed that nearly half the number of interviewees were unable to describe an identifiable procedure or process which they may implement in their own EAP testing experience. Whilst it may be the case that interviewees were unable to define a process due to the need to vary approach based on the context of testing or assessment (Calderhead & Shorrock, 1997, p.194), the researcher raises the concern that the situation may also be attributable to a gap between theory and practice. Consequently, this situation could be addressed through the provision of additional support which highlights the good practice associated with following a template (Davidson & Lynch, 2002) or structured approach to assessment design and usage.

The concerns addressed in the section immediately above can also be linked to the results of items AQ16B and the findings from AQ17. AQ16B shows that 64.5% of respondents recognised the need to use a test specification (Ruch, 1924) in the process of EAP test design. AQ17 demonstrated that the stage of test design associated with reference to or creation of a test specification was amongst the most frequently omitted items.

6.2.1 Observations relating to views on/experience of validity and reliability in EAP testing and assessment

With reference to aspects of construct validity, the Likert scale results based on items drawn from AQ16 and AQ19 show a varied pattern of opinion regarding the importance of domain specificity in EAP assessment (Carroll, 1981, p.67, Kane, 2012, p.41-42; Weir, 1983). Although only 41.2% of respondents directly support the creation of assessments designed to represent specific domains of academic study. As discussed in the analysis section, this situation may reflect common dichotomies associated with ESAP and EGAP in the field. Despite this spread of views, 65% of respondents to AQ16h did report a need for different marking descriptors in different contexts. This difference can potentially be attributable to the fact that contexts differ within subjects as well as across subject areas leading to a more immediate requirement to consider particular types of assessment tasks rather than the need to consider subject specialism. This theory may also be seen to align with Fulcher's (1999) view that an excessive concern for specificity can distract from the wider scope of construct validity.

In response to a Likert scale drawn from AQ20, it does also appear that the majority of teachers recognise benefits in researching the features of different subject areas which students are studying. In addition, benefits are seen in working more closely with academic colleagues from different subject fields, with regard to the potential impact on enhancing validity of EAP testing and assessment. The benefits of working with other academic colleagues across the institution, in order to understand the features of academic subject areas, were also supported by DVAS item AQ16C. Furthermore, the results of item AQ22G also show that practitioners have a strong sense of responsibility to assist other colleagues in understanding test scores. Groups of comments from the interview stage also support the benefits to validity of collaborating with subject teachers across the institution. This can be seen from responses to BQ10. Collaboration for the purposes of

moderation was also referred to in a group of 7 comments in response to AQ17. One concern is however acknowledged in BQ10 where comments are also made regarding potential barriers to collaboration with colleagues within institutions. This reference to the key importance of collaboration in the process of EAP assessment, supported by testing researchers such as Davidson and lynch (2002) and Jafarpur (2003, p.72), appears to be an area of activity which is widely supported by EAP teachers involved in assessment. In the view of this researcher, it should therefore be promoted and encouraged even more widely so that optimal benefits can be obtained.

Finally, with regard to reliability, it has been noted, through respondents' comments in AQ19B in the questionnaire, AQ24 from the free text comments and interview items BQ13 and BQ17, that concerns have been expressed regarding the reliability of IELTS. The researcher has then raised further concerns associated with response to AQ22J which reveals that 55.6% of respondents are either unsure or in agreement with the creation of such IELTS equivalences based on in-house tests. The researcher considers these findings disturbing, given that both the test specifications and grading criteria for the standardised tests mentioned are not commonly accessible in the public domain and difficulties may arise in relation to criterion referencing.

6.2.2 Observations relating to practices involving analysis and interpretation of the results and scores of EAP tests and assessments

Further evidence of EAP teachers' observance of principles associated with Assessment Literacy can be identified in multiple references to outcomes-based approaches. This is evident in comments relating to the provision of feedback drawn from AQ22. This is then supported with similar comments made in individual items such as AQ16H and interview responses to BQ7, BQ9, and BQ11And BQ13.

Observance of this trend can be considered to show a high frequency of concern for familiarity with assessing and interpreting the achievement in focus in a principled and consistent manner.This is one of the key aspects of Assessment literacy as identified by Stiggins (1995).

Likert scale items within the questionnaire referring to the use of statistics in the interpretation of test scores revealed a clear indication that the majority of respondents are not sufficiently familiar with either descriptive statistics or inferential statistics in the process of analysing tests and test scores. Whilst overall familiarity with descriptive statistics was higher than that of inferential statistics, those with experience of using either set of procedures was notably low. The lack of familiarity with these analytical tools suggests the need for a form of training that can demonstrate the powerful contribution which statistics can make, in a manner which is accessible to the non-statistically-minded EAP teacher.

6.2.3 Observations relating to practices involving the consideration of ethics in EAP testing and assessment

With regard to ethics, an important principle for Assessment Literacy, the results of the Likert scale items from AQ25, the free-text items from AQ27 and the comments made in BQ17 and 18 from the interview show that EAP teachers involved in assessment tend to adopt an ethical stance which favours the needs and best interests of test-takers. This is paralleled by recommendations for good practice issued by champions of Assessment Literacy and critical language testing (McNamara & Roever, 2006; Popham, 2001; Shohamy, 2001; Stiggins, 1995). Nevertheless, when respondents were asked about the specific situations associated with their own practice through the Likert scale related to AQ26, only 41% felt able to declare that their employment and assessment practice was not associated with unethical practice. The researcher therefore

recommends a more widespread campaign for awareness raising, related to ethics in EAP testing and assessment, undertaken by EAP teachers.Importantly, this should involve the wider institutions in which EAP teachers are located, in order that the dangers of unethical practice associated with EAP assessment can be understood by all the necessary stakeholders.

6.3 Consideration of conclusions related to research questions and the hypothesis

In order to determine how the conclusions from this study relate to the objectives and original drivers for this research project, it is crucial at this stage to consider both the research questions which steered the research and the hypothesis which the research sought to reject or prove.

The first research question for the study asked:

> 'To what extent do EAP teacher views on EAP testing and assessment practices reflect language testing research and practices which comprise Assessment Literacy?'

As referred to above, a number of areas of good practice associated with Assessment Literacy in EAP, which can be linked to theory from research, have been identified in the practice of EAP teachers. The conclusions also demonstrate that there also appear to be key areas where the assessment practices of EAP teachers can be improved or enhanced significantly with closer reference to theories emerging from language testing research and Assessment Literacy. The second research question asks, 'How can EAP Assessment Literacy be sustained or enhanced?'

In response to this question, suggestions which provide a series of possible answers linked to the conclusions are provided in the Recommendations below in Section 6.5.

Finally, in order to consider the hypothesis behind this research the null hypothesis (H_o) has been tested in order to determine if it can be rejected thus providing support for the original hypothesis. The null hypothesis is as follows:

EAP teachers who are involved in EAP testing and assessment do not have any identifiable development requirements with regard to their knowledge and ability to implement assessment good practice and recommendations stemming from research.

Based on the conclusions cited in the section above, it is concluded that the null hypothesis can be rejected and that the original hypothesis has been supported.

6.4 Limitations

Despite the observations which have been described, which the researcher suggests can be generalised to the wider sector of EAP assessment, there are a number of limitations to this study which need to be considered.

Firstly, it should be noted that the observations arrived at through analysis of data collected through the research tools used in this study represent the opinions of EAP teachers regarding their own Assessment literacy rather than a more objective measure of Assessment Literacy undertaken through observation of assessment in practice. With regard to the questionnaire, the number of questionnaire respondents was restricted to practitioners with membership of professional bodies, which may have excluded less engaged practitioners who have not sought membership of such an organisation. In addition, whilst the number of respondents was not small, a greater number of respondents may have influenced the nature of the data gathered. Regarding

the questionnaire tool itself, it is clear to the researcher that additional time devoted to the piloting stage could have resulted in the honing of certain items. This may have led to larger clusters of items being suitable for use in composite Likert scales. It should also be noted that responses to free-text items were smaller in volume as these items were marked as optional.

As far as the interview stage is concerned, whilst the data collected was rich, the number interviewees conducted was restricted to 25 and therefore more limited in scope than the questionnaire. Had there been time and sufficient resources, a higher number of interviewees may have yielded additional and richer information. With respect to the interactional profiles, it may also be the case that had a trained interviewer been employed the interactions would have been of a different nature and the information revealed by interviewees could have been even more accurate. With regard to the content analysis of the interviews, whilst the views of colleagues were consulted in order to corroborate and criticise components or factors which were identified, the consultation with a larger group of colleagues may have helped identify other trends.

6.5 Recommendations and concluding comments

In order to summarise the recommendations made in this chapter, which have emerged from the conclusions cited in the section above, a series of five key action points has been identified as a means of drawing together the next steps which should follow on from this research. It is the researcher's belief that this series of recommendations, when considered in relation to the conclusions which have been drawn, represent a unique and focused understanding of the challenge of developing and maintaining Assessment Literacy amongst EAP teachers. It also offers a practical and realistic set of tools for the realisation of practical initiatives for the enhancement of Assessment Literacy in this context.

It is the researcher's aim to create a series of internet-based resources which will provide a user-friendly and accessible mechanism for the implementation of the recommendations listed below. It is also hoped that the suggested initiatives for maintaining and sustaining Assessment Literacy which are provided, may also act as a stimulus for further research by practitioners across the sector. To reiterate a reoccurring theme which is characteristic of Assessment Literacy and its supporters, such as Popham (2001) and Stiggins (1995), the success of the mission to enhance and maintain high standards of and expanded Assessment Literacy (Taylor, 2009) in the context of EAP lies in the proactivity of the practitioners themselves and thus the actioning of recommendations such as those provided at this final point of thesis.

6.5.1 Key action points for the enhancement and maintenance of EAP teacher Assessment Literacy stemming from this thesis:

6.5.1.1 Provide a downloadable and transferable framework for in-service EAP Assessment Literacy training

The researcher intends to devise and provide a web-based transferable framework for the design and implementation of in-service training interventions by EAP departments which will assist in building practitioner skill in areas associated with EAP assessment. This resource could also draw attention to the key role played by in-service skill development through the complex yet necessary challenges of EAP assessment. Importantly the resource would allow departments to identify the EAP assessment literacy training needs which are specific to their staff. In this way such activity would take the form of a series of group-based self-help initiatives which could be implemented according to local needs and resources in different EAP departments. It is hoped that through this

mechanism confidence and interest in EAP assessment can be built according to local requirements and simultatenously key aspects of Assessment Literacy can be fostered. In terms of materials, this may comprise a series of reflective exercises and practical activities designed according to key aspects of Assessment Literacy, as outlined in the Literature Review. As an example, this could include practice exercises in the development and use of test specifications, one key stage of EAP test development which this research has shown to be frequently overlooked. An exercise of this type could highlight the benefits of principled EAP test creation associated with following a template or a structured approach to assessment design. Another example of a possible training intervention in this series could include an exercise which focuses on skills associated with analytical or interpretive stages of EAP test development and usage. This might involve the development and use of grading criteria and the use of methods of statistical analysis.

6.5.1.2 Create an online annotated bibliography to foster research-led EAP Assessment Literacy

The creation of a medium for proactively introducing practitioners to published research is advocated. Ideally, this should be achieved by making EAP teachers aware of research which is relevant to EAP Assessment Literacy, in a manner which is perceived to be accessible. In this manner, knowledge embedded within these resources could be more easily absorbed and disseminated. This objective could be achieved through the provison of an online annotated bibliography which would allow EAP practitioners to identify key research materials relavant to EAP Assessment Literacy. A brief and straightforward summary of ways in which certain publications and research papers might be best applied to testing practice, would also be included.

6.5.1.3 Promote collaborative research and learning opportunities in EAP assessment, both within institutions and across the sector

The researcher intends to promote collaborative research and learning opportunities, linked to EAP assessment Literacy, both within and beyond individual institutions. Such opportunities could include cross-institutional groups or pairings which can foster action research into EAP assessment. More structured and collective opportunities for research into EAP assessment and testing could also be instigated. Means could also be sought to promote and develop research opportunities and interactions between EAP practitioners and academic subject specialists within EAP practitioners' home institutions. This would assist in better operationalising relevant constructs in EAP tests and assessments and could enhance the validity of inferences drawn from test scores.

6.5.1.4 Develop interpretive EAP Assessment Literacy skills through statistical analysis training, with the use of online video tutorials

The researcher intends to explore mechanisms to develop EAP teachers' interpretive Assessment Literacy skills through statistical analysis. It is hoped that this will be beneficial for EAP teachers, who are not already skilled in this area and who are not naturally motivated to develop these skills. The intention is to achieve this aim through creating a series of video tutorials, provided online, showing how a software package, such as SPSS can be used to explore and analyse tests results. Whilst a number of books are already available which provide guidance on the use of statistics to analyse language tests, these are arguably still complex to follow. As a result, video tutorials could

be utilised in order to demonstrate statistical procedures in a lighter manner which could be designed to appear less technical and more straightforward.

6.5.1.5 Raise awareness about the key importance of ethicality as an aspect of EAP Assessment Literacy, through an online forum

An online forum can be provided to raise awareness amongst EAP teachers and their institutions, regarding the dangers of unethical practice associated with EAP assessment. This could involve the more prominent publication of and adherence to a code of ethics, related to EAP Assessment, such as that published by EALTA (Erickson & Figueras, 2010). EAP practitioners could also be encouraged to discuss ethical concerns more proactively through an online forum tool. Mechanisms could also be considered to educate colleagues across educational institutions beyond the confines of EAP departments, regarding the unreliability of standardised tests if their scores are interpreted inappropriately. This could involve briefing sessions for central admissions departments regarding the features and scope of standardised EAP tests and how results should be appropriately interpreted. This process might also draw attention to the dangers associated with the unprincipled use of estimated equivalences to scores from standardized tests, if these have been generated unscientifically.

CHAPTER 7 REFERENCES

AALL. (2012). Association for Academic Language and Learning Retrieved 16 July, 2012, from http://www.aall.org.au/

Adams, P. (2008). Considering 'best practice': The social construction of teacher activity and pupil learning as performance Cambridge Journal of Education 38(3), 375-392.

Adler, P. A., & Adler, P. (1994). Observational techniques. In N. K. Denzin & Y. S. Lincoln (Eds.), Handbook of Qualitative Research. London: Sage.

Alderson, J. C. (1988). New procedures for validating proficiency tests of ESP? Theory and Practice. Language Testing(5), 2.

Alderson, J. C. (2000). Assessing reading. Cambridge: Cambridge University Press.

Alderson, J. C., Clapham, C., & Wall, D. (1995). Language test construction and evaluation. Cambridge: Cambridge University Press.

Alderson, J. C., & Hughes, A. (1981). Issues in language testing. London: British Council.

Alexander, O., Argent, S., & Spencer, J. (2008). EAP essentials : a teacher's guide to principles and practice. Reading: Garnet.

Allison, D. (1999). Language testing and evaluation : an introductory course. Singapore ; River Edge, N.J.: Singapore University Press ;World Scientific.

Allwright, D., & Bailey, K. M. (1991). Focus on the language classroom: an introduction to classroom research for language teachers. Cambridge: Cambridge University Press.

Allwright, J. (1997). The induction programme for the 4-week pre-sessional course at Lancaster University. Paper presented at the BALEAP Professional Interest Reports -1995-1997Meetings, London.

ALTE (Producer). (2012, 4, November 2012). About ALTE. Retrieved from http://www.alte.org/about_alte

Anderson, L. W., & Krathwohl, D. (Eds.). (2001). A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives. New York: Longman

Ansell, M. A. (2008). The privatisation of English for Academic Purposes teaching in British universities. Liaison, 1 (July 2008), 18-19.

Anstey, E. (1966). Psychological testing. London: Nelson.

Apple. (2009). Ipod Classic 5th generation. Retrieved 31 August, 2009, from http://www.apple.com/ipodclassic/

ASKe. (2013). Assessment Standards Knowledge Exchange, 2013, from http://www.brookes.ac.uk/aske/

Assessment Reform Group. (2002). Assessment for learning: 10 principles Retrieved from http://www.uni-koeln.de/hf/konstrukt/didaktik/benotung/assessment_basis.pdf

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

Bachman, L. F. (2000). Modern language testing at the turn of the century: assuring that what we count counts. Language Testing, 17(1), 1-42.

Bachman, L. F. (2004). Statistical analyses for language assessment. Cambridge: Cambridge University Press.

Bachman, L. F. & Cohen, A. D. (1998) (Eds.). Interfaces between second language acquisition and language testing research (pp. xiv, 202p.). Cambridge: Cambridge University Press.

Bachman, L. F., & Kunnan, A. J. (2005). Statistical analyses for language assessment: workbook and CD-ROM. Cambridge: Cambridge University Press.

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

Bailey, K. M., & Brown, J. D. (1996). Language testing courses: What are they? In A. Cumming & R. Berwick (Eds.), Validation in language testing Clevedon, UK: Multilingual Matters.

BALEAP. (2008). TEAP Working Party: The Competency Framework for Teachers of English for Academic Purposes 2011, from http://www.baleap.org.uk

BALEAP. (2012a). BALEAP Working Party on Testing: BALEAP Guidelines on English Language tests for university entry 2012, from http://www.baleap.org.uk/projects/testing-working-party

BALEAP. (2012b). The Global Forum for EAP Professionals Retrieved 16 July 2012, 2012, from http://www.baleap.org.uk/home

BALEAP. (2013). Tracking EAP students working party. from http://www.baleap.org.uk

Banerjee, J., & Wall, D. (2006). Assessing and reporting performances on pre-sessional EAP courses: Developing a final assessment checklist and investigating its validity. Journal of English for Academic Purposes, 5(1), 50-69.

Benesch, S. (2001). Critical English for academic purposes : theory, politics, and practice. Mahwah, N.J. ; London: Lawrence Erlbaum Associates.

Berman, M. (1998). A multiple intelligences road to an ELT classroom. London: Crown House.

Berry, R. (2008). Assessment for learning. Hong Kong London: Hong Kong University Press ;Eurospan [distributor].

Bethlehem, J. G. (2009). Applied survey methods: a statistical perspective. Oxford: Wiley-Blackwell.

Biber, D., Conrad, S. M., Reppen, R., Byrd, P., Helt, M., Clark, V., Urzua, A. (2004). Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus Monograph Series, 25.

Biggs, J. B., & Tang, C. S. (2011). Teaching for quality learning at university: what the student does (4th ed. ed.). Maidenhead: McGraw-Hill/Society for Research into Higher Education/Open University Press.

Black, P. J. (2003). Assessment for learning : putting it into practice. Maidenhead: Open University Press.

Blaxter, L., Hughes, C., & Tight, M. (2006). How to research (3rd ed.). Maidenhead: Open University Press.

Bloom, B. S. (1956). Taxonomy of educational objectives: The classification of educational goals, by a committee of college and university examiners. Handbook 1: Cognitive domain. New York: Longmans.

Blue, G. M., Milton, J., & Saville, J. (2000). Assessing English for academic purposes. Oxford: New York: P. Lang.

Bogdan, R., & Biklen, S. K. (1992). Qualitative research for education: an introduction to theory and methods (2nd ed.). Boston ; London: Allyn and Bacon.

Brindley, G. (2001). Language assessment and professional development Studies in Language Testing - Experimenting with uncertainty: Essays in honour of Alan Davies (Vol. 11, pp. 126-133). Cambridge: Cambridge University Press.

Bristol University. (Producer). (2012, 11 November 2012). Bristol Online Surveys. Retrieved from http://www.survey.bris.ac.uk/

Brookes, A., & Grundy, P. D. (1990). Writing for study purposes : a teacher's guide to developing individual writing skills. Cambridge: Cambridge University Press.

Brown, G. T. L. (2008). Assessment Literacy Training and Teachers' Conceptions of Assessment. New York: Nova Science.

Brown, H. D., & Abeywickrama, P. (2010). Language assessment : principles and classroom practices. White Plains, NY: Pearson/Longman.

Brown, J. D. (1988). Understanding research in second language learning : a teacher's guide to statistics and research design. Cambridge: Cambridge University Press.

Brown, J. D. (2001). Using surveys in language programs. Cambridge: Cambridge University Press.

Brown, J. D., & Bailey, K. M. (2008). Language testing courses: What are they in 2007? Language Testing, 25(3), 349-383.

Bruce, I. (2011). The centrality of genre in EAP instruction and research: Addressing issues of theoretical diversity and construct validity. Paper presented at the BALEAP, University of Portsmouth.

Bryman, A. (2008). Social research methods (3rd ed. ed.). Oxford: Oxford University Press.

Bryman, A. (2012). Social research methods (4th ed. ed.). Oxford: Oxford University Press.

Burton, S. (2012, 12 April 2012). From where I sit - Time's not on our side. Times Higher Education.

Bynner, J., & Stribley, K. M. (1979). Social research: principles and procedures. London: Longman, in association with Open University Press.

Byram, M. (2004). Routledge encyclopedia of language teaching and learning. London: Routledge.

Calderhead, J., & Shorrock, S. B. (1997). Understanding teacher education. London: Falmer.

Cambridge, U. o. (Producer). (2012a). Cambridge language assessment. English Language Teaching. Retrieved from http://www.cambridge.org/gb/elt/catalogue/subject/item382381/Cambridge-Language-Assessment/?site_locale=en_GB

Cambridge University. (Producer). (2012b). Studies in Language Testing. Cambridge English: Research & Validation. Retrieved from http://research.cambridgeesol.org/research-collaboration/silt

Cardinal, D., Hayward, J., & Jones, G. (2004). Epistemology: the theory of knowledge: London : John Murray, 2004 (2005 printing).

Carrell, P. L., Devine, J., & Eskey, D. E. (1988). Interactive approaches to second language reading. Cambridge: Cambridge University Press.

Carroll, B. J. (1981). Specifications for an English Language Testing Service. In J. C. Alderson & A. Hughes (Eds.), Issues in Language Testing. Oxford: Pergamon Press.

Center for Advanced Research on Language Acquisition (CARLA). (2012, 30 Aptil 2012). Virtual Assessment Center (VAC) Retrieved June 2012, 2012, from http://www.carla.umn.edu/assessment/vac/

Center for Applied Linguistics. (2007). Foreign Language Assessment Directory. http://www.cal.org/CALWebDB/FLAD

Center for Applied Linguistics. (2008). Understanding Assessment: A Guide for Foreign

Language Educators. http://www.cal.org/flad/tutorial/index.html

Chalhoub-Deville, M., Chapelle, C. A., & Duff, P. (2006). Inference and generalizability in applied linguistics : multiple perspectives. Amsterdam: John Benjamins.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), Interfaces between second language acquisition and language testing research (pp. xiv, 202p.). Cambridge: Cambridge University Press.

Chapelle, C. A. (1999). Validity in language assessment. Annual Review of Applied Linguistics(19), 254-272.

Chapelle, C., & Douglas, D. (2006). Assessing language through computer technology. Cambridge: Cambridge University Press.

Cizek, G. J. (1995). The big picture in assessment and who ought to have it. The Phi Delta Kappan, 77(3), 246-249.

Clapham, C. (1996). The development of IELTS: a study of the effect of background knowledge on reading comprehension. Cambridge: Cambridge University Press.

Cohen, L., Manion, L., & Morrison, K. (2007). Research methods in education (6th ed.). London: Routledge.

Coniam, D. (2009). Investigating the quality of teacher-produced tests for EFL students and the effects of training in test development principles and practices on improving test quality. System, 37(2), 226.

Cortina, B. M. (1993). What is coeffieicent alpha?An examination of theory and applications.". Journal of Applied Psychology(78), 98-104.

Council of Europe. (2003). Relating language Examinations to the Common European Framework of Reference for languages:Learning, Teaching, Assessment (CEF): Manual-Prelininary Pilot Version. Strasbourg: Council of Europe Language Policy Division.

Cowie, A. P., & Heaton, J. B. (1977). English for Academic Purposes : papers on the language problems of overseas students in Higher Education in the UK. Reading: British Association for Applied Linguistics.

Creswell, J. W. (2003). Research design : qualitative, quantitative, and mixed methods approaches (2nd ed. ed.). Thousand Oaks, Calif. ; London: Sage.

Cronbach, L. J. (1990). Essentials of psychological testing (5th ed.). New York ; London: Harper & Row.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin(52), 281-302.

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

Cudeck, R. (2001). Measurement. Journal of Consumer Psychology, 10(1&2), 55-69.

Daniels, H. (1993). Charting the agenda : educational activity after Vygotsky. London: Routledge.

Datta, L. (1994). Paradigm wars:A basis for peaceful co-existence and beyond. In C. S. Reichardt & S. F. Rallis (Eds.), The Qualitative Quantitative Debate: New perspectives (Vol. 61, pp. 85-91). San Francisco, CA: Jossey-Bass.

Davidson, F., & Lynch, B. K. (2002). Testcraft : a teacher's guide to writing and using language test specifications. New Haven ; London: Yale University Press.

Davies, A. (1968). Language testing symposium: a psycholinguistic approach; edited by Alan Davies. London: Oxford University Press.

Davies, A. (1990). Principles of language testing. Oxford: Basil Blackwell.

Davies, A. (1997). Demands of being professional in language testing. Language Testing, 14(3), 328-339.

Davies, A., Brown A., C Elder, C., Hill, K., Lumley, T. & McNamara, T. (1999) Dictionary of Language Testing. Cambridge: Cambridge University Press.

Davies, A., & Elder, C. (2004). The handbook of applied linguistics. Malden, MA ; Oxford: Blackwell.

DCSF. (2008). The Assessment for Learning Strategy Retrieved from https://www.education.gov.uk/publications/eOrderingDownload/DCSF-00341-2008.pdf

De Lotbinière, M. (2011, 11 January, 2011). UK seeks to raise language entry levels, The Guardian. Retrieved from http://www.guardian.co.uk/education/2011/jan/11/tefl-internationaleducationnews1?INTCMP=SRCH

De Vaus, D. A. (2002). Surveys in social research (5th ed. ed.). London: UCL Press.

Denscombe, M. (2003). The good research guide for small-scale social research projects (2nd ed. ed.). Maidenhead: Open University Press.

Denscombe, M. (2007). The good research guide for small-scale social research projects (3rd ed. ed.). Maidenhead: Open University Press.

Denzin, N. K., & Lincoln, Y. S. (2005). The Sage handbook of qualitative research (3rd ed.). Thousand Oaks, Calif. ; London: Sage.

Ding, Y., & Foo, S. (2002). Ontology research and development. Part 2 - a review of ontology mapping and evolving. Journal of Information Science, 28(5), 375-388.

Doff, A. (1988). Teach English: a training course for teachers. Cambridge: Cambridge University Press in association with the British Council.

Doff, A. (1995). Teach English: a training course for teachers. Cambridge: Cambridge University Press.

Dörnyei, Z. (2003). Questionnaires in second language research: construction, administration, and processing (2nd ed. ed.). New York: Routledge.

Dörnyei, Z. (2007). Research methods in applied linguistics: quantitative, qualitative, and mixed methodologies. Oxford: Oxford University Press.

Douglas, D. (2000). Assessing languages for specific purposes. Cambridge: Cambridge University Press.

Drever, E. (2003). Using semi-structured interviews in small-scale research: a teacher's guide. Glasgow: Scottish Council for Research in Education.

Drever, E., & Scottish Council for Research in Education. (2003). Using semi-structured interviews in small-scale research : a teacher's guide (Rev. ed.). Glasgow: Scottish Council for Research in Education.

Driscoll, A., & Wood, S. (2007). Developing Outcomes-Based Assessment for Learner-Centered Education: A faculty introduction. Sterling, Virginia: Stylus Publishing.

Dudley-Evans, T., & St John, M. J. (1998). Developments in ESP : a multi-disciplinary approach. Cambridge: Cambridge University Press.

Duff, P. (2002). Research approaches in applied linguistics. In R. B. Kaplan (Ed.), The Oxford handbook of applied linguistics (pp. xxvii, 641 p.). Oxford: Oxford University Press.

Duranti, A. (1997). Linguistic anthropology. Cambridge: Cambridge University Press.

EAQUALS. (2012). Background to the CEFR Retrieved 30/07, 2012, from http://www.eaquals.org/pages/?p=7043

Erickson, G., & Figueras, N. (2010). EALTA Guidelines for Good Practice in Language Testing and Assessment: Large scale dissemination days, from http://www.ealta.eu.org

ETS. (2011). TOEFL Program History. Insight- TOEFL IBT Research, 6(1).

Evers, J. C. F. (2011). From the Past into the Future. How Technological Developments Change Our Ways of Data Collection, Transcription and Analysis Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, (12(1)). Retrieved from http://nbnresolving.de/urn:nbn:de:0114-fqs1101381.

Feyerabend, P. (1970). Consolations for the Secialist. In I. E. Lakatos & A. E. Musgrave (Eds.), Criticism and the growth of knowledge : proceedings of the International Colloquium in the Philosophy of Science, London, 1965, volume 4. Cambridge: Cambridge University

Fink, A. (2008). Practicing research: discovering evidence that matters. Los Angeles, Calif. ; London: SAGE.

Flowerdew, J., & Peacock, M. (2001). Research perspectives on English for academic purposes. Cambridge: Cambridge University Press.

Forehand, M. (2005). Bloom's taxonomy. Emerging perspectives on learning, teaching, and technology. Retrieved 23 July, 2012, from http://projects.coe.uga.edu/epltt/index.php?title=Bloom%27s_Taxonomy

Foucault, M. (1977). Discipline and punish : the birth of the prison. London: Penguin.

Foucault, M., & Gordon, C. (1980). Power/knowledge : selected interviews and other writings, 1972/1977. Brighton: Harvester Press.

Fry, H., Ketteridge, S., & Marshall, S. (1999). A handbook for teaching and learning in higher education : enhancing academic practice. London: Kogan Page.

Fry, H., Ketteridge, S., & Marshall, S. (2003). A handbook for teaching and learning in higher education : enhancing academic practice (2nd ed. ed.). London: Kogan Page.

Fulcher, G. (1999). Assessment in English for academic purposes: putting content validity in its place. Applied Linguistics, 20(2), 221-236.

Fulcher, G. (2006). Test architecture. Foreign Language Education Research(9), 1-22.

Fulcher, G. (2007, April 13). Universities Undermine Their Own Foundations: Contracting out English preparation courses in Britain is a short-term fix, The Guardian.

Fulcher, G. (2009a). The Commercialisation of Language Provision at University. In J. C. Alderson (Ed.), The Politics of Language Education: Individuals and institutions (pp. 125-143). Bristol: Multilingual Matters.

Fulcher, G. (2009b). Language Testing for Higher Education. Language Testing Scenarios - Language Testing Resources Retrieved 4 August 2012, 2012, from http://languagetesting.info/whatis/scenarios/4hedu.php

Fulcher, G. (2009c). Language Testing Resources Retrieved 4 August 2012, 2012, from http://languagetesting.info/

Fulcher, G. (2010). Practical language testing. London: Hodder Education.

Fulcher, G. (2012a). Assessment literacy for the language classroom. [Journal Article]. Language Assessment Quarterly, 9(1).

Fulcher, G. (2012b). The Routledge Handbook of Language Testing Oxford: Routledge.

Fulcher, G., & Davidson, F. (2007). Language testing and assessment: an advanced resource book. London: Routledge.

Fulcher, G., & Davidson, F. (2009). Test architecture, test retrofit. Language Testing, 26(1), 123-144.

Fuller, S. (2006). Kuhn vs Popper : the struggle for the soul of science. Thriplow: Icon.

Gage, N. L. (1989). The Paradigm Wars and Their Aftermath A "Historical" Sketch of Research on Teaching Since 1989. Educational Researcher, 18(7), 4-10.

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

Giddens, A. (2001). Sociology (4th ed.). Cambridge: Polity.

Gillham, B. (2000a). Case study research methods. London: Continuum.

Gillham, B. (2000b). Developing a questionnaire. London: Continuum.

Gillham, B. (2000c). The research interview. London: Continuum Publishing.

Gillham, B. (2005). Research interviewing : the range of techniques. Maidenhead: Open University Press.

Gipps, C. (1994). Beyond testing : towards a theory of educational assessment. London: Falmer.

Gorard, S. (2001). Quantitative methods in educational research : the role of numbers made easy. London: Continuum.

Gottlieb, M. H., & Nguyen, D. (2007). Assessment and accountability in language education programs : a guide for administrators and teachers. Philadelphia, PA: Caslon Pub.

Green, A. (2007). IELTS washback in context : preparation for academic writing in higher education. Cambridge: Cambridge University Press.

Greene, J. C. (2008). Mixed methods in social inquiry. San Francisco, Calif.: Jossey-Bass ; Chichester : John Wiley [distributor].

Griffin Technology. (2009). iTalk Pro Retrieved 31 August 2009, from http://www.griffintechnology.com/products/italkpro

Grove, J. (2012). London Met banned from enrolling overseas students Times Higher Education.

Gruber, T. (1993). A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 5(2), 199-220.

Guba, E. G., & Lincoln, Y. S. (1989). Fourth generation evaluation. Newbury Park, Calif. ; London: Sage.

Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research The Handbook of Qualitative Research (pp. 105-117). Thousand Oaks, CA: Sage.

Guba, E. G., & Lincoln, Y. S. (2005). Controversies, Contradictions, Confluences. In N. K. Denzin & Y. S. Lincoln (Eds.), The Sage handbook of qualitative research (3rd ed., pp. 200-201). Thousand Oaks, Calif.; London: Sage.

Gubrium, J. F., & Holstein, J. A. (2002). Handbook of interview research : context & method. Thousand Oaks; London: Sage Publications.

Guse, J., & Thornbury, S. (2011). Communicative activities for EAP. Cambridge ; New York: Cambridge University Press.

Gwartney, P. A. (2007). The telephone interviewer's handbook : how to conduct standardized conversations. San Francisco, Calif.: Jossey-Bass ; Chichester : John Wiley [distributor].

Hamp-Lyons, L. (2011). English for academic purposes: 2011 and beyond. Journal of English for Academic Purposes, 10(1), 2-4.

Harmer, J. (2007). The practice of English language teaching (4th ed.). Harlow: Longman.

Healey, J. F. (2012). Statistics : a tool for social research (9th ed.). Australia ; United Kingdom: Wadsworth, Cengage Learning.

Heigham, J., & Croker, R. A. (2009). Qualitative research in applied linguistics: a practical introduction. Basingstoke: Palgrave Macmillan.

Henning, G. (1987). A guide to language testing: development, evaluation, research. Cambridge, MA: Newberry House Publishers.

Henning, G. (1988). The influence of test and sample dimensionality on latent trait person ability and item difficulty calibrations. Language Testing, 1(5), 83-89.

Heywood, J. (2000). Assessment in higher education: student learning, teaching, programmes and institutions. London: Jessica Kingsley.

Hinkel, E. (2005). Handbook of research in second language teaching and learning. Mahwah, N.J. ; London: Lawrence Erlbaum Associates.

Hinton, P. R. (2004). SPSS explained. London: Routledge.

Hughes, A. (1988). Testing English for university study. [Oxford]: Modern English Publications in association with the British Council.

Hughes, A. (2003). Testing English for University Study. ELT Documents 127. Oxford: Modern English Press.

Hutchison,T.,& Waters, A. (1985). ESP at the crossroads. In Swales, J. (Ed.), Episodes in ESP: a source and reference book on the development of English for Science and Technology. Oxford: Pergamon Institute of English.

Hyland, K. (2006). English for Academic Purposes - and Advanced resource book: Routledge.

Hyland, K., & Hamp-Lyons, L. (2002). EAP: issues and directions. Journal of English for Academic Purposes, 1(1), 1-12.

IELTS. (2008). IELTS History Retrieved 27 March 2008, from http://www.ielts.org/aboutus/article284.aspx

IELTS. (2012). The History of IELTS Retrieved 4, August, 2012, from http://www.ielts.org/researchers/history_of_ielts.aspx

ILTA. (2000). ILTA Code of Ethics Retrieved 19 November 2008, 2008, from http://www.iltaonline.com/

ILTA. (2012). Annual LTRC Conferences Retrieved 30/07, 2012, from http://www.iltaonline.com/index.php?option=com_content&view=article&id=132&Itemid=144

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. Language Testing, 25(3), 385-402.

Jafarpur, A. (2003). Is the test constructor a facet? Language Testing, 20(1), 57-87.

Jordan, R. R. (1997). English for academic purposes : a guide and resource book for teachers. Cambridge: Cambridge University Press.

Jordan, R. R. (2002). The growth of EAP in Britain. Journal of English for Academic Purposes, 1, 69-78.

Kane, M. (2012). Articulating a validity argument. In F. Davidson & G. Fulcher (Eds.), The Routledge Handbook of Language Testing (pp. 34-47). Oxford: Routledge.

Kaplan, A. (1973). The conduct of inquiry. Methodology for behavioral science: Aylesbury: Intertext.

Kerlinger, F. N., & Lee, H. B. (2000). Foundations of behavioral research (4th ed. / Fred N. Kerlinger, Howard B. Lee. ed.). Fort Worth ; London: Harcourt College Publishers.

Kincheloe, J. L., & McLaren, P. (2005). Rethinking Crtical Teoy and Qualitative Research The Sage handbook of qualitative research (3rd ed., pp. xix, 1210 p.). Thousand Oaks, Calif.; London: Sage.

Kirk, R. E. (2008). Statistics : an introduction (5th ed. ed.). Belmont: Thomson Wadsworth.

Knight, P. (1995). Assessment for learning in higher education. London: Kogan Page.

Krueger, R. A. (2000). Focus groups : a practical guide for applied research (3rd ed.). Thousand Oaks, Calif.; London: Sage.

Krzanowski, M. (2001). 'S/He Holds the Trinity/UCLES Dip: Are They Ready to Teach EAP?'. Paper presented at the BALEAP PIM: Teacher Training for EAP, University of Bath. http://www.baleap.org.uk/baleap/conference-events/pims/pim-reports/teacher-training-eap/

Kuhn, T. S. (1970). The structure of scientific revolutions. Second edition, enlarged: [Chicago, London:] University of Chicago Press.

Kunan, A. J. (2003). Test fairness. Paper presented at the European Year of Language Conference, Barcelona.

Kyodo (Producer). (2012, January 15). University entrance exams kick off. The Japan Times. Retrieved from http://www.japantimes.co.jp/text/nn20120115b1.html

Kyriacou, C. (1991). Essential teaching skills. Oxford: Basil Blackwell.

Kyriacou, C. (2007). Essential teaching skills (3rd ed. ed.). Cheltenham: Nelson Thornes.

Lado, R. (1961). Language testing: the construction and use of foreign language tests, a teacher's book: Longmans.

Larsen-Freeman, D., & Long, M. H. (1991). An introduction to second language acquisition research. London: Longman.

Larson-Hall, J. (2010). A guide to doing statistics in second language research using SPSS. London: Routledge.

Lather, P. (1992). Critical Frames in Educational Research: Feminist and Post-Structural Perspectives. Theory into Practice, 31(2), 87.

Law, S. (2007). Philosophy. London: Dorling Kindersley.

Lerner, G. H. (2004). Conversation analysis: studies from the first generation. Amsterdam: John Benjamins Publishing.

Lewkowicz, J. A. (2000). Authenticity in language testing: some outstanding questions. Language Testing, 17(1), 385-402.

Lichtman, M. (2006). Qualitative research in education: a user's guide. Thousand Oaks, Calif. ; London: SAGE.

Linn, R. L., National Council on Measurement in Education., & American Council on Education. (1988). Educational measurement (3rd ed.). New York: American Council on Education.

Lipsett, A. (2008, November 12). Newcastle expels overseas students with fake grades, The Guardian. Retrieved from http://www.guardian.co.uk/education/2008/nov/12/fake-grades-students-china?INTCMP=SRCH

LTF (Producer). (2012, 4, November, 2012). 32nd Language Testing forum 2012 Retrieved from http://www.bristol.ac.uk/education/ltf2012/

Lukin, L. E., Bandalos, D. L., Eckhout, T. J., & Mickelson, K. (2004). Facilitating the Development of Assessment Literacy. Educational Measurement: Issues and Practice, 23(2), 26-32.

Lynch, B. K. (1997). In search of the ethical test. Language Testing, 14(3), 315-327.

Lynch, B. K. (2001). Rethinking assessment from from a critical perspective. Language Testing, 18(4), 351-372.

Lynch, B. K., & Shaw, P. (2005). Portfolios, power and ethics. TESOL Quarterly, 39(2), 263-297.

Malone, M. E. (2011). Assessment Literacy for Language Educators. CAL Digest(Ocober 2011).

Masden, H. S. (1983). Techniques in Testing Oxford Oxford University Press.

Mason, J. (2002). Qualitative researching (2nd ed.). London: Sage.

Mason, R. (2012, 12 July, 2012). Vince Cable: efforts to cut student immigration 'damaging' to UK science, The Telegraph. Retrieved from http://www.telegraph.co.uk/journalists/rowena-mason/9396025/Vince-Cable-efforts-to-cut-student-immigration-damaging-to-UK-science.html

Master, P. (2005). Research in English for Specific Purposes. In E. Hinkel (Ed.), Handbook of research in second language teaching and learning (pp. 101-102). Mahwah, N.J. ; London: Lawrence Erlbaum Associates.

McLaren, P., & Giarelli, J. M. (1995). Critical theory and educational research. Albany, N.Y.: State University of New York Press.

McNamara, T. F. (1996). Measuring second language performance. London: Longman.

McNamara, T. F. (2000). Language testing. Oxford: Oxford University Press.

McNamara, T. F. (2001). Language assessment as social practice: challenges for research. Language Testing, 18(4), 333-349.

McNamara, T. F. (2003). Tearing us apart again: The paradigm wars and the search for validity Eurosla Yearbook, 3, 229-238.

McNamara, T.F, & Roever, C. (2006). Language testing: the social dimension. Oxford: Blackwell.

Meikle, J., & Malik, S. (2012, 30, August 2012). London Met crisis will damage UK's brand, says vice-chancellor, The Guardian. Retrieved from http://www.guardian.co.uk/education/2012/aug/30/london-metropolitan-university-international-students

Menter, I. (2010). Literature review on teacher education in the 21st Century Retrieved from http://www.scotland.gov.uk/Publications/2010/09/23094515/0

Mertens, D. M. (1998). Research methods in education and psychology : integrating diversity with quantitative & qualitative approaches. Thousand Oaks, Calif London: Sage Publications.

Mertens, D. M. (2003). Mixed methods and the politics of human research- The transformative-emancipatory perspective Handbook of mixed methods in social & behavioral research (pp. xv, 768 p.). Thousand Oaks, Calif. ; London: SAGE.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. [610p.].). New York: American Council on Education.

Miles, M. B., & Huberman, A. M. (1994). Qualitative data analysis: an expanded sourcebook (2nd ed.). Thousand Oaks, Calif. ; London: Sage.

Moore, T., & Morton, J. (2005). Dimensions of difference: a comparison of university writing and IELTS writing. Journal of English for Academic Purposes, 4(1), 43-66.

Moreno Espinosa, S. (2003). Vocabulary: Reviewing Trends in EFL/ESL Instruction and Testing. ODISEA: Journal of English Studies, 4, 97-112. Retrieved from http://www.ual.es/odisea/Odisea04_MorenoEspinosa.pdf

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in Educational Assessment. Review of Research in Education, 30(ArticleType: research-article / Issue Title: Special Issue on Rethinking Learning: What Counts as Learning and What Learning Counts / Full publication date: 2006 / Copyright © 2006 American Educational Research Association), 109-162.

Munby, J. (1978). Communicative syllabus design: a sociolinguistic model for defining the content of purpose-specific language programmes. Cambridge: Cambridge University Press.

NCH Software. (2009). Express Scribe Transcription Playback Software Retrieved 31 August, 2009

Nesfield-Cookson, B. (1987). William Blake : prophet of universal brotherhood. Wellingborough: Crucible.

Northumbria University. (2011). Guidelines for Good Assessment Practice at Northumbria University. In L. a. t. support (Ed.), (pp. 9). Newcastle Upon Tyne. Retrieved from www.northumbria.ac.uk/static/worddocuments/ardocs/ggap.doc.

Nunan, D. (1992). Research methods in language learning. Cambridge: Cambridge University Press.

O'Sullivan, B. (2007). Validation of in-house testing. Paper presented at the BALEAP PIM: Testing and Assessment in EAP, Salford.

O'Sullivan, B. (Eds) (2011). Language testing: theories and practices. Basingstoke: Palgrave Macmillan.

Oller, J. W. (1979). Language tests at school: a pragmatic approach. London: Longman.

Phillips. (2009). Foot Control Retrieved 31 August, 2009, from http://www.dictation.philips.com/index.php?id=1667&CC=SE&uid=202

Popham, W. J. (1971). Criterion-referenced measurement: an introduction. Englewood Cliffs, N.J.: Educational Technology Publications.

Popham, W. J. (2001). The truth about testing : an educator's call to action. Alexandria, Va.: Association for Supervision and Curriculum Development.

Popham, W. J. (Producer). (2006). Needed: a dose of assessment literacy. Educational Leadership. Retrieved from http://www.ascd.org/publications/educational-leadership/mar06/vol63/num06/Needed@-A-Dose-of-Assessment-Literacy.aspx

Popham, W. J. (2009). Assessment literacy for teachers: faddish or fundamental? Theory into practice(48), 4-11.

Popham, W. J. (2012a). Assessing Students' Affect (2nd ed.): Pearson.

Popham, W. J. (2012b). Mastering assessment: a self-service system for educators. New York ; Abingdon: Routledge.

Popham, W. J. (2012c). Validity: Assessment's cornerstone (2nd ed.): Pearson.

Potter, J., & Hepburn, A. (2009). Transcription Retrieved 31 August, 2009, from http://www-staff.lboro.ac.uk/~ssjap/transcription/transcription.htm

Price, M., Rust, C., O'Donovan, B., & Handley, K. (2012). Assessment Literacy: The foundation for improving student learning. Oxford: Oxford Brookes University.

Raimes, A. (1991). Out of the Woods: Emerging Traditions in the Teaching of Writing TESOL Quarterly, 25(3), 407-430

Rea-Dickins, P., Kiely, R. & Yu, G. (2011). Uses and impact of test scores in university admissions processes: the language test as the ‘hard’ criterion. in: O’Sullivan, B. (Eds) (2011) Language Testing: Theories and Practices. Palgrave, 262-281.

Reichardt, C. S., & Rallis, S. F. (1994). Qualitative and quantitative inquiries are not incompatible:A call for a new partnership. . New Directions for Program Evaluation 61, 85-91.

Richardson, L., & Adams St.Pierre, E. (2005). Writing- a Method of Inquiry The Sage handbook of qualitative research (3rd ed., pp. xix, 1210 p.). Thousand Oaks, Calif. ; London: Sage.

Roberts, P. (2001). Teacher Training for EAP. Paper presented at the BALEAP PIM: Teacher Training for EAP, University of Bath. http://www.baleap.org.uk/baleap/conference-events/pims/pim-reports/teacher-training-eap/

Robinson, P. C. (1980). ESP (English for Specific Purposes) : the present position. Oxford: Pergamon.

Robson, C. (2002a). Real world research: a resource for social scientists and practitioner-researchers (2nd ed. ed.). Madden, Mass. ; Oxford: Blackwell Publishers.

Robson, C. (2002b). Real world research : a resource for social scientists and practitioner-researchers (2nd ed.). Oxford: Blackwell.

Rubin, H. J., & Rubin, I. (2005). Qualitative interviewing: the art of hearing data (2nd ed.). Thousand Oaks, Calif. ; London: Sage.

Ruch, G. M. (1924). The Improvement of the Written Examination. New York: Scott, Foresman and Company.

Sauro, J. (Producer). (2011, 11 August 2012). Are both positive and negative items necessary in questionnaires? Measuring Usaability: Quantitative Usanility, Statistics & Six Sigma. Retrieved from http://www.measuringusability.com/positive-negative.php

Sauro, J., & Lewis, J. (2011, 7-12 May 2011). When Designing Usability Questionnaires, Does It Hurt to Be Positive? Paper presented at the CHI 2011, Vancouver, BC, Canada.

Schmitt, D. (2012). EAP Assessment in the UK. Paper presented at the BALEAP AGM 2012, London South Bank University. http://www.baleap.org.uk/projects/teap-working-party

Schwandt, T. A. (1994). Constructivist, interpretivist approaches to human inquiry. Handbook of qualitative research. Thousand Oaks, Calif. ; London: Sage.

Scott, D., & Usher, R. (1996). Understanding educational research. London: Routledge.

Scrivener, J. (2011). Learning teaching : the essential guide to English language teaching (3rd ed. ed.). Oxford: Macmillan Education.

Seliger, H. W., & Shohamy, E. (1989). Second language research methods. Oxford: Oxford University Press.

Selinker, L., Tarone, E., & Hanzeli, V. E. (1981). English for academic and technical purposes : studies in honor of Louis Trimble. Rowley, Massachusetts: Newbury House.

Sharpling, G., P. (2002). Learning to teach English for Academic Purposes: Some current training and development issues. English Language Teacher Education and Development (ELTED), 6.

Sharpling, G., P. (2007). Testing writing for academic purposes: seven common dilemmas and seven possible solutions. Paper presented at the BALEAP PIM: Testing and Assessment in EAP, Salford.

Sharpling, G., P. (2008). The development of a dedicated English language entry test for foundation programme students: the case of WELT. InForm(1), 7-9.

Sharpling, G., P. (2010). WELT Handbook Retrieved 9, August, 2012, from http://www2.warwick.ac.uk/fac/soc/al/centre/centrespecialisms/testing/tests/welt/handbook

Sharpling, G., P., & Sky, G. (2010). Introduction to special issue: Teacher development, testing and assessment English Language Teacher Education and Development (ELTED), 13.

Sheldon, L. E. (2004). Directions for the future : issues in English for academic purposes. Bern ; Oxford: Peter Lang.

Shohamy, E. (1998). Critical language testing and beyond. Studies in Educational Evaluation, 24(4), 331-345.

Shohamy, E. (2001). The power of tests: a critical perspective on the uses of language tests. Harlow: Longman.

SIG, I. E. (2012). IATEFL: English for Specific Purposes Retrieved 16 July, 2012, from http://espsig.iatefl.org/index.html

Smith, J. K. (1983). Quantitative versus qualitative research: an attempt to clarify the issue. Educational Researcher, 12(3), 6-13.

Smith, M. L. (1994). Qualitative plus/versus quantitative: The last word. In C. S. E. Reichardt & S. F. E. Rallis (Eds.), The qualitative-quantitative debate: new perspectives : Plenary sessions : Annual conference : Papers. San Francisco, CA: Jossey-Bass.

Spack, R. (1988). Initiating ESL students into academic discourse community: how far should we go? TESOL Quarterly, 22(1), 29-52.

Spolsky, B. (1976). Language Testing: Art of Science? Paper presented at the 4th International Congress of Applied Linguistics, Stuttgart, Germany.

Spolsky, B. (1981). Some ethical questions about language testing. I. In C. Klein-Braley & D. K. Stevenson (Eds.), Practice and problems in language testing I (pp. 5-21). Frankfurt Am Main: Lang.

Spolsky, B. (1995). Measured words: the development of objective language testing. Oxford: Oxford University Press.

Spolsky, B. (2012). Language testing and language management. In F. Davidson & G. Fulcher (Eds.), The Routledge Handbook of Language Testing (pp. 497). Oxford: Routledge.

Stansfield, M. (2001). Introduction to Paradigms. Victoria, Canada: Trafford.

Stern, H. H. (1983). Fundamental concepts of language teaching. Oxford: Oxford University Press.

Stern, H. H. (1992). Issues and options in language teaching. Oxford: Oxford University Press.

Stiggins, R. J. (1991). Assessment Literacy. The Phi Delta Kappan, 72(7), 534.

Stiggins, R. J. (1995). Assessment Literacy for the 21st Century. The Phi Delta Kappan, 77(3), 238.

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. The Phi Delta Kappan, 83(10), 758-765.

Stobart, G. (2008). Testing times: the uses and abuses of assessment. London: Routledge.

Swain, H. (2012, 7 May, 2012). Patrick McGhee, new chair of Million+ thinktank, is trending nationally, The Guardian. Retrieved from http://www.guardian.co.uk/education/2012/may/07/patrick-mcghee-chair-millionplus-thinktank?INTCMP=SRCH

Swales, J. (1985). Episodes in ESP: a source and reference book on the development of English for Science and Technology. Oxford: Pergamon Institute of English.

Tashakkori, A., & Teddlie, C. (1998). Mixed methodology: combining qualitative and quantitative approaches. Thousand Oaks, Calif. ; London: Sage.

Tashakkori, A., & Teddlie, C. (2003). Handbook of mixed methods in social & behavioral research. Thousand Oaks, Calif. ; London: SAGE.

Taylor, L. (2009). Developing Assessment Literacy. Annual review of Applied Linguistics, 29, 21-36.

TESOL. (2012). TESOL International Association Retrieved 16 July, 2012, from http://www.tesol.org/

Tuckman, B. W. (1972). Conducting educational research.

Tuckman, B. W. (1999). Conducting educational research (5th ed. ed.). Fort Worth: Harcourt Brace College Publishers.

UCU. (2012). Fighting Privatisation in Tertiary Education: Campaigns against INTO Retrieved 17 July 2012, 2012, from http://www.ucu.org.uk/stopprivatisation

UKBA. (December 2010). The Student Immigration System: A Consultation. Home Office Retrieved from http://www.ukba.homeoffice.gov.uk/sitecontent/documents/policyandlaw/consultations/students/

University of Reading. (2004). TEEP Examination Report 2004Ur, P. (1999). A course in language teaching : trainee book. Cambridge: Cambridge University Press.

Ur, P. (1999). A course in English language teaching (1st ed.). Cambridge: Cambridge University Press.

Ur, P. (2012). A course in English language teaching (2nd ed.). Cambridge: Cambridge University Press.

Urmson, J. O., & Rée, J. (2005). The concise encyclopedia of western philosophy (3rd ed. ed.). London: Routledge.

UUK. (2010). Universities UK response to comments by the Immigration Minister Retrieved 10 August 2012, 2010, from http://www.universitiesuk.ac.uk/Newsroom/Media-Releases/Pages/UniversitiesUKresponsetocommentsbytheImmigrationMinister.aspx

Van der Walt, J. L., & Faans, S. J. (2008). The validation of language tests. Stellenbosch Papers in Linguistics, 38, 191-204.

Van Lier, L. (1988). The classroom and the language learner : ethnography and second-language classroom research. London: Longman.

Von Dietze, E. (2001). Paradigms explained : rethinking Thomas Kuhn's philosophy of science. Westport, Connecticut ; London: Praeger.

Walker, R., Adelman, C., & Wiedel, J. (1975). A guide to classroom observation. London: Methuen.

Wallace, C. (1997). IELTS: global implications of curriculum and materials design. ELT J, 51(4), 370-373.

Wallace, M. J. (1991). Training foreign language teachers : a reflective approach. Cambridge: Cambridge University Press.

Walters, F. S. (2011). Cultivating Assessment Literacy: Standards Evaluation Through Language-Test Specification Reverse Engineering. Language Assessment Quarterly, 7(4), 317.

Warren, C. A. B. (2002). Qualitative Interviewing. In J. F. Gubrium & J. A. Holstein (Eds.), Handbook of interview research. Thousand Oaks, Calif. ; London: Sage Publications.

Waters, A. (1996). A Review of Research into Needs in English for Academic Purposes of Relevance to the North American Higher Education Context Monograph Series, 33.

Weigle, S. C. (2002). Assessing writing. Cambridge: Cambridge University Press.

Weir, C. J. (1983). Identifying the language problems of overseas students in tertiary education in the United Kingdom. Thesis PhD. London:University of London Institute of Education.

Weir, C. 1988. The specification, realization and validation of an English language proficiency test. In A. Hughes (Ed.) Testing English for University Study. ELT Documents 127. Oxford: Modern English Press.

Weir, C. J. (1990). Communicative language testing. New York: Prentice Hall.

Weir, C. J. (1993). Understanding and developing language tests. New York ; London: Prentice Hall.

Weir, C.J. (2005). Language Testing and Validation. Basingstoke: Palgrave Macmillan.

Weston, P. (1989). Turning them off or bringing them on?: assessment motivation and learning. Paper presented at the Assessment and learning: papers presented at the members conference, Slough.

Widdowson, H. G. (1983). Learning purpose and language use. Oxford: Oxford University Press.

Wigdor, A. K., & Garner, W. R. (1982). Ability testing: uses, consequences and controversies. Washington, D.C.: National Academy Press.

Wigglesworth, G., & Elder, C. (1996). Perspectives on the testing cycle: Setting the scene. Australian Review of Applied Linguistics Series S(13), 13-32.

Woodhouse, M. B. (1996). Paradigm wars : worldviews for a New Age. Berkeley, Calif.: Frog.

Woods, A., Fletcher, P., & Hughes, A. (1986). Statistics in language studies. Cambridge: Cambridge University Press.

Wooffitt, R., & Widdicombe, S. (2006). Interaction in interviews. In P. Drew, G. Raymond & D. Weinberg (Eds.), Talk and interaction in social research methods (pp. xi, 233 p.). London: Sage.

Wright, F. (2012, April 2012). EAP in the UK – how to achieve a life-changing transformation, EL Gazette.

Yin, R. K. (2003). Case study research : design and methods (3rd ed.). Thousand Oaks, Calif. ; London: Sage.

CHAPTER 8 APPENDIX

8.1 Full Questionnaire

Page 1 of 10

Survey on EAP Tutor Testing & Assessment Practices

Page 1 of 10 - Welcome

This survey is designed to collect your views on testing and assessment, as a professional in the field of English for Academic Purposes (EAP). The results of the survey will be used as part of a doctoral research project on **Applied Linguistics**, conducted through the **University of Leicester**.

At the end of the survey all participants will be able to opt into a prize draw for **£50** worth of Amazon E-vouchers with at least a 1 in 30 chance of winning.

The main survey questions are on **pages 3-8** and should take you **around 30 minutes** to complete. Please note that the terms 'Testing' and 'Assessment' are used interchangeably in this survey. Please base your responses on any experience you have in any form of EAP assessment or testing, whether it be large or small scale, classroom or examination based.

Once you click **'continue'** below, you will be directed to an initial section:

Page 2 - Data Protection and Ethics.

The main survey will then start on **Page 3**. It is not possible to return to a page once it has been completed. The main sections of the survey are organised as follows:

Page 3- Personal details and EAP education and assessment involvement
Page 4- Training and skills in EAP testing & assessment
Page 5- EAP test and assessment design processes
Page 6- Validity and reliability in EAP testing and assessment
Page 7- Analysis and interpretation of EAP tests & assessment results
Page 8- Ethical considerations in EAP testing and assessment

The final two pages of the survey are structured as follows:

Page 9- Invitations to join a prize draw and a subsequent interview stage
Page 10- Thank you and confirmation of data submission

Many thanks for your participation.

Anthony Manning (EdD Student- University of Leicester)

Page 2 of 10 - Data protection & ethics

Data Protection

All data collected in this survey will be held anonymously and securely. Should you feel able to provide your email address for contact for the interview stage at the end of the survey, neither your identity nor your place of work will be disclosed or linked to the information which you have provided.

1. I agree that my data, gathered through this survey, may be stored (after anonymisation) in the researcher's records and may be used for the purpose of the research project described.

○ Yes ○ No

Ethics

To read more about the nature of the research being conducted in this project, the Research Information Sheet for this project is available below:

Download Research Information Sheet

If you have any further questions please feel free to email me, Anthony Manning, on am360@leicester.ac.uk

2. I confirm that I have been provided with the **Research Information Sheet** for the above study and have been given the opportunity to ask further questions.

○ Yes ○ No

3. I understand that my participation is voluntary and that should I wish to withdraw at any point, I am able to do so without explanation.

○ Yes ○ No

4. I agree to take part in this survey which is part of the above study described on **Page 1**.

○ Yes ○ No

5. I agree that the information I provide can be used by the researcher in anonymised form.

○ Yes ○ No

Page 3 of 10 - Personal details and EAP education and assessment involvement

Your personal details

6. I am:

○ Male ○ Female

7. My age bracket is:

Select an answer

8. I consider my nationality to be most closely associated with the following country:

Select an answer

If you selected Other, please specify:

Your EAP education and assessment involvement

9. Please answer 'Yes', 'Unsure' or 'No' to the following statements:

	Select only one answer per row		
	Yes	**Unsure**	**No**
a. My career has involved some experience as a teacher/tutor/lecturer in the field of **English for Academic Purposes** (EAP).	○	○	○
b. I have some experience in creating EAP-related tests and/or assessments.	○	○	○
c. I have some experience of marking EAP tests and/or assessments.	○	○	○

d. Some of the tests or assessments that I have worked on have an impact on whether students can progress onto other courses of academic study.			

10. The country where I have most recently been involved with EAP testing and assessment is:

Select an answer

If you selected Other, please specify:

11. The most recent institution where I have been involved with EAP testing and assessment, can best be described as a(n):

Select an answer

If you selected Other, please specify:

12. I first worked in the field of EAP:

Select an answer

Page 4 of 10 - Your training and skills in EAP testing & assessment

Your training and skills in EAP testing & assessment

13. Please select the option which best describes your attitude towards the following statements:

	Select only one answer per row				
	Strongly agree	**Agree**	**Neither agree nor disagree**	**Disagree**	**Strongly disagree**
a. My interest in EAP is driven by aspects of the field other than					

testing or assessment.					
b. Testing and assessment skills are crucial for my role as an EAP tutor.	○	○	○	○	○
c. The skills that I have acquired in EAP assessment have mainly been developed through courses I have taken which include focus on language testing.	○	○	○	○	○
d. I regularly read books or articles on language testing.	○	○	○	○	○
e. EAP testing and assessment is one of my key skill areas.	○	○	○	○	○
f. The reference and guidance material available to me on the subject of language testing and assessment is very user-friendly.	○	○	○	○	○
g. I feel that I have had sufficient training in EAP assessment and testing practices.	○	○	○	○	○
h. It is more straightforward to test skills in EAP than to test subject knowledge.	○	○	○	○	○
i. It is straightforward to identify the key EAP language areas and skills that students need to be assessed in.	○	○	○	○	○
j. It is straightforward to recognize a question which is not working properly in an EAP test or assessment.	○	○	○	○	○

14. Please rate your skills in the following areas

	Select only one answer per row				
	Very highly skilled	**Highly skilled**	**Average skilled**	**Low skilled**	**Very low skilled**

a. Test and assessment design					
b. Defining test purpose					
c. Identifying what should be assessed					
d. Trialling/piloting tests					
e. Writing test items/questions					
f. Developing and using grading procedures					
g. Using descriptive statistics (such as mean, mode and median) to analyse test and assessment scores					
h. Using inferential statistics (such as Cronbach's Alpha, T-tests and factor analyses) to analyse test and assessment scores					
i. Drawing inferences from test and assessment scores					
j. Making instructional decisions based on test scores					
k. Giving feedback to students based on test scores					
l. Helping other people who use EAP test results to understand them					
m. Identifying problems with tests or assessments					

15. Please add any additional comment relating to your **training and skills in EAP testing & assessment**. You can also use this section if you wish to add any comment to clarify responses given above. *(Optional)*

Page 5 of 10 - Your views on EAP test and assessment design processes

Your views on EAP test and assessment design processes

16. Please select the option which best describes your attitude towards the following statements:

	Select only one answer per row				
	Strongly agree	**Agree**	**Neither agree nor disagree**	**Disagree**	**Strongly disagree**
a. With practice, EAP test design can become a straightforward process.					
b. It is necessary to use a test specification when designing a test.					
c. It is unhelpful to involve academic content tutors in the process of EAP test design.					
d. When designing an EAP test or assessment, it is necessary to take into account the specific subjects that students will be studying.					
e. Each EAP testing situation will require its own bespoke test.					
f. A good starting point in EAP testing is to write the test questions.					
g. EAP test design is best left to the individual class teacher.					
h. Each different testing and assessment situation may need its own marking descriptors.					

i. It would be appropriate to design an EAP diagnostic test, in order to measure which students can use English effectively for university study.					
j. An EAP achievement test would be a suitable test to design in order to evaluate a learner's understanding of a specific EAP skill.					

17. Please select an option from each of the 8 drop-down lists below in order to suggest your preferred sequence for the **essential** stages of EAP test design. For this exercise you are restricted to 8 stages but have more than 8 options.

	Select a different option for each of the 8 stages.
a. Stage 1	Select an answer
b. Stage 2	Select an answer
c. Stage 3	Select an answer
d. Stage 4	Select an answer
e. Stage 5	Select an answer
f. Stage 6	Select an answer
g. Stage 7	Select an answer
h. Stage 8	Select an answer

18. Please add any additional comment relating to your views and experience of **EAP test and assessment design**. You can also use this section if you wish to add any comment to clarify responses given above. *(Optional)*

Your views on validity and reliability in EAP testing and assessment

19. Please select the option which best describes your attitude towards the following statements:

	Select only one answer per row				
	Strongly agree	**Agree**	**Neither agree nor disagree**	**Disagree**	**Strongly disagree**
a. It is difficult to be specific about what constitutes a particular area of linguistic skill in EAP.	○	○	○	○	○
b. Decisions based on test results from IELTS or TOEFL are more likely to be valid and reliable than those based on the results of my own EAP assessments.	○	○	○	○	○
c. EAP tests should use reading and listening texts created specifically for EAP language testing.	○	○	○	○	○
d. When creating EAP tests, it is unreasonable to expect EAP tutors to use language from the different academic fields in which students' English will be applied.	○	○	○	○	○
e. When selecting EAP writing exercises and assessments for students preparing to study Mathematics, it would be best to avoid essays.	○	○	○	○	○
f. Students studying different academic subjects are likely to need different types of EAP assessment.	○	○	○	○	○
g. When a test is well compiled, a single question can be an effective way to measure any given language ability.	○	○	○	○	○

h. A reliable EAP test will also be valid.					
i. Giving students a different version of one of my EAP tests, aimed at assessing the same skills, would likely result in students getting a similar mark.					
j. A class of students gaining a low grade in an EAP assessment is a good indication that the assessment mechanism was faulty in some way.					

20. In your opinion, are the actions below likely to enhance validity?

	Select only one answer per row		
	Yes	**Unsure**	**No**
a. Researching the subject contexts in which students are required to use EAP			
b. Incorporating multiple measures of EAP language abilities			
c. Adding more multiple choice EAP questions			
d. Requiring all EAP tutors in a team to mark all test sections			
e. Working as a team with other EAP colleagues to determine what should be tested			
f. Allowing EAP markers to use their own judgement when grading			
g. Phrasing EAP questions so that they sound academic			
h. Including a question for each area of EAP which has been taught in the course			

21. Please add any additional comment relating to your views on **validity and reliability** in EAP testing & assessment. You can also use this section if you wish to add any comment to clarify responses given above. *(Optional)*

Page 7 of 10 - Your views on analysis & interpretation of EAP tests & assessment results

Your views on analysis and interpretation of EAP tests & assessment results

22. Please select the option which best describes your attitude towards the following statements:

	Select only one answer per row				
	Strongly agree	**Agree**	**Neither agree nor disagree**	**Disagree**	**Strongly disagree**
a. Analysing the results of EAP tests and assessments is a relatively straightforward process.					
b. It is really difficult to draw concrete conclusions from EAP test and assessment results alone.					
c. The results from one good EAP test can generally provide a very accurate overall impression of a student's breadth of ability.					
d. EAP students who are required to reach a certain level of ability in their EAP skills should be evaluated in relation to specific learning outcomes (criterion referencing).					
e. The results of EAP tests should be explained in more depth so that other colleagues can understand					

what skills they relate to.					
f. I am frequently required to help academics in other departments to understand students' EAP grades more effectively.	○	○	○	○	○
g. The EAP department is responsible for making sure that users of the test results interpret their meaning appropriately.	○	○	○	○	○
h. A class of 17 students all gain low scores in section 3 of an EAP test. This is most likely to mean that the students need to develop their skills in that area further.	○	○	○	○	○
i. EAP students who need to demonstrate a particular skill level ought to be assessed and selected based on their rank compared to other students who completed the test (norm referencing).	○	○	○	○	○
j. It is impractical to convert EAP test results into IELTS or TOEFL grades as the process is too complex.	○	○	○	○	○

23. How familiar are you with the following statistical terms and procedures?

	Select only one answer per row				
	Very familiar (I have used this myself)	**Familiar (but I have not used this myself)**	**Neither familiar nor unfamiliar**	**Unfamiliar**	**Very unfamiliar**
a. Normal distribution	○	○	○	○	○
b. Standard deviation	○	○	○	○	○
c. Factor analysis	○	○	○	○	○

d. Mean					
e. Cronbach's Alpha					
f. Median					
g. Split half analysis					
h. Mode					
i. T-tests					
j. Scatter plots					

24. Please add any additional comment relating to your views on **analysis and interpretation of EAP test and assessment results**. You can also use this section if you wish to add any comment to clarify responses given above *(Optional)*

Page 8 of 10 - Your views on ethical considerations in EAP testing and assessment

Your views on ethical considerations in EAP testing and assessment

25. Please select the option which best describes your attitude towards the following statements:

	Select only one answer per row				
	Strongly agree	**Agree**	**Neither agree nor disagree**	**Disagree**	**Strongly disagree**
a. Chinese students struggle with EAP speaking tests and should be required to gain higher grades in EAP speaking tests from next year.					
b. The University has asked to use the Pre-sessional EAP test with a					

group of nurses seeking to find employment at a local hospital. The Academic Language Centre should allow this.					
c. A security risk has required the evacuation of the building for 15 minutes during group A's EAP test, as a result the tests should be marked as they were at the point of evacuation.					
d. Two tutors from the EAP team would like to give extra classes to certain students who they feel need to improve their examination technique. This action should be avoided					
e. A University department wants to recruit fewer international students, so the Academic Language Centre should permit the change of the passing grade for current students from 60 to 70%.					
f. Government policy has changed, so it is acceptable to tell a current class of EAP students that they either need to improve their admissions scores by 10%, or return home.					
g. Unexpectedly, nearly all students fail the final EAP test so the EAP team should consider adjusting the marks using a bell curve.					
h. An EAP tutor is preparing students for an EAP test. It is therefore good practice for the tutor to highlight key skills for revision in the class course textbook.					
i. University rules require students to show their library card at the door of the Exam room. Wen Hu					

has lost his library card so he should not be allowed to enter the room.					
j. Students should be encouraged to question the grades they get in EAP tests.	○	○	○	○	○

26. Please answer 'Yes', 'Unsure' or 'No' to the following statements:

	Select only one answer per row		
	Yes	**Unsure**	**No**
a. I am familiar with codes of practice for language testing.	○	○	◉
b. I am concerned about the social implications of EAP language testing.	○	○	○
c. Language testers should make it clear that they have all the relevant knowledge regarding language assessment.	○	○	○
d. I am responsible for evaluating the ethical consequences of the projects I'm involved with.	○	○	○
e. I feel able to refuse to participate in procedures which would violate ethical principles of language testing.	○	○	○
f. I feel that my employers are sufficiently aware of ethical dilemmas associated with language testing.	○	○	○
g. I have recently felt obliged to contribute to EAP testing procedures which I question the ethicality of.	○	○	○
h. New regulations connected to language qualification and visa issuance have been introduced in	○	○	○

an ethical manner.			

27. Please add any additional comment relating to your views on **ethical considerations in EAP testing & assessment**. You can also use this section if you wish to add any comment to clarify responses given above. *(Optional)*

9 of 10 - Invitations to join a prize draw and a subsequent interview stage

Prize draw

28. If you would like to be entered into a prize draw with the opportunity to win £50 worth of Amazon vouchers, please write your email address in the box below: *(Optional)*

Interview participation

29. If you are willing to be contacted by email, to arrange a short telephone or Skype interview of no more than 20 minutes in duration, please write your email address in the box below: *(Optional)*

Page 10 of 10 - Thank you and confirmation of data submission

Thank you. Your data has been submitted and this is the end of the survey.

If you have any further questions please feel free to email am360@le.ac.uk. If you have given your email address for prize either draw or the interview stage you will be contacted in due course.

8.2 Interview Questions

Introductory Information

- Thank you very much for agreeing to speak to me today. I really appreciate your giving me this time. It shouldn't take more than about 15 minutes.
- The reason why I am conducting these interviews is to examine tutors practices in EAP assessment, and testing. As you know, I'm doing this as part of my studies for an EdD via the University of Leicester.
- Here is a copy of the research information sheet for further details (pass Research Information Sheet)
- As you would expect, I won't link any data to any individuals, so you'll remain totally anonymous, as will the institution that you work for. The information you provide will be safeguarded in line with data protection regulations.
- So that I can properly analyse the information that you provide, I would like to record this conversation, are you happy for me to do that?
- OK then; if you're ready, I'll begin, please just ask me to repeat if anything isn't clear.

BQ1	Before we begin, I'd like to give you the opportunity to tell me any thoughts you may have on the topic of EAP testing and Assessment, in case you have anything to say before I ask any specific questions. Do you have any thoughts on the subject of EAP language assessment, before we begin?
BQ2	How would you say you have developed or acquired skills in EAP assessment and testing during your career and education?
BQ3	What would you say are your strengths and weaknesses in EAP testing and assessment?
BQ4	What do you find most challenging about EAP testing and assessment?
BQ5	How do you normally go about designing an EAP test or assessment? Do you follow a set process when you create a new EAP test or assessment? If so what is it?
BQ6	How do you go decide which questions to put in or leave out in EAP tests and assessments?
BQ7	How do you know that your tests are valid?
BQ8	How do you know that your tests are reliable?
BQ9	How do you define the domain of EAP which forms the basis of your tests or assessments?

BQ10	How do you measure or interpret the results of EAP tests or assessments?
BQ11	How do you know that your interpretations are accurate?
BQ12	How do you help test takers and, where appropriate, other users of test scores understand the results of tests and assessments that you create?
BQ13	To what extent is your testing and assessment influenced by language testing research? If so, in which ways?
BQ14	Do you use any particular books or journals which help with EAP testing and assessment?
BQ15	Are there any other resources that assist you with EAP testing and research?
BQ16	Do you have any concerns about the activity of EAP testing and assessment?
BQ17	Would you say that EAP testing requires any different considerations to other forms of English language testing?
BQ18	Is there anything else that you would like to say about the topic of EAP assessment?

8.3 Transcripts of Interviews

Interview Code	Web Link to download Interview Transcripts
Interview A	http://dl.dropbox.com/u/6563593/Interview%20Transcripts%20-%20Anthony%20Manning/A-Interview-Transcript.docx
Interview B	http://dl.dropbox.com/u/6563593/Interview%20Transcripts%20-%20Anthony%20Manning/B-Interview-Transcript.docx
Interview C	http://dl.dropbox.com/u/6563593/Interview%20Transcripts%20-%20Anthony%20Manning/C-Interview-Transcript.docx
Interview D	http://dl.dropbox.com/u/6563593/Interview%20Transcripts%20-%20Anthony%20Manning/D-Interview-Transcript.docx
Interview E	http://dl.dropbox.com/u/6563593/Interview%20Transcripts%20-%20Anthony%20Manning/E-Interview-Transcript.doc
Interview F	http://dl.dropbox.com/u/6563593/Interview%20Transcripts%20-%20Anthony%20Manning/F-Interview-Transcript.doc
Interview G	http://dl.dropbox.com/u/6563593/Interview%20Transcripts%20-%20Anthony%20Manning/G-Interview-Transcript.doc
Interview H	http://dl.dropbox.com/u/6563593/Interview%20Transcripts%20-%20Anthony%20Manning/H-Interview-Transcript.doc
Interview I	http://dl.dropbox.com/u/6563593/Interview%20Transcripts%20-%20Anthony%20Manning/I-Interview-Transcript.docx
Interview J	http://dl.dropbox.com/u/6563593/Interview%20Transcripts%20-%20Anthony%20Manning/J-Interview-Transcript.doc
Interview K	http://dl.dropbox.com/u/6563593/Interview%20Transcripts%20-%20Anthony%20Manning/K-Interview-Transcript.doc
Interview L	http://dl.dropbox.com/u/6563593/Interview%20Transcripts%20-%20Anthony%20Manning/L-Interview-Transcript.doc
Interview M	http://dl.dropbox.com/u/6563593/Interview%20Transcripts%20-%20Anthony%20Manning/M-Interview-Transcript.doc
Interview N	http://dl.dropbox.com/u/6563593/Interview%20Transcripts%20-%20Anthony%20Manning/N-Interview-Transcript.doc

Interview O	http://dl.dropbox.com/u/6563593/Interview%20Transcripts%20-%20Anthony%20Manning/O-Interview-Transcript.doc
Interview P	http://dl.dropbox.com/u/6563593/Interview%20Transcripts%20-%20Anthony%20Manning/P-Interview-Transcript.doc
Interview Q	http://dl.dropbox.com/u/6563593/Interview%20Transcripts%20-%20Anthony%20Manning/Q-Interview-Transcript.doc
Interview R	http://dl.dropbox.com/u/6563593/Interview%20Transcripts%20-%20Anthony%20Manning/R-Interview-Transcript.doc
Interview S	http://dl.dropbox.com/u/6563593/Interview%20Transcripts%20-%20Anthony%20Manning/S-Interview-Transcript.doc
Interview T	http://dl.dropbox.com/u/6563593/Interview%20Transcripts%20-%20Anthony%20Manning/T-Interview-Transcript.doc
Interview U	http://dl.dropbox.com/u/6563593/Interview%20Transcripts%20-%20Anthony%20Manning/U-Interview-Transcript.doc
Interview V	http://dl.dropbox.com/u/6563593/Interview%20Transcripts%20-%20Anthony%20Manning/V-Interview-Transcript.docx
Interview W	http://dl.dropbox.com/u/6563593/Interview%20Transcripts%20-%20Anthony%20Manning/W-Interview-Transcript.doc
Interview X	http://dl.dropbox.com/u/6563593/Interview%20Transcripts%20-%20Anthony%20Manning/X-Interview-Transcript.doc
Interview Y	http://dl.dropbox.com/u/6563593/Interview%20Transcripts%20-%20Anthony%20Manning/Y-Interview-Transcript.doc

8.4 Responses to open questions from questionnaire

The responses to open questions from the interview can be downloaded here:

Open Question Responses – Overseas Respondents	http://dl.dropbox.com/u/6563593/Questionnaire%20Open%20Questions/Open%20Questions%20Questionnaire%20-%20Overseas.docx
Open Question Responses UK Respondents	http://dl.dropbox.com/u/6563593/Questionnaire%20Open%20Questions/Open%20Questions%20Questionnaire-UK.docx

8.5 Interview Analysis Grid

The grid used for Interview content analysis can be downloaded here:

Interview Analysis Grid	http://dl.dropbox.com/u/6563593/Interview%20Content%20Analysis%20Grid/Interview%20Content%20Analysis%20Grid.xlsx

8.6 Full Chi Square outputs for Findings set A1 (158 population)

(C-S1) Respondent gender and country of most recent involvement with EAP assessment

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
6. I am: * 10. The country where I have most recently been involved with EAP testing and assessment is:	158	100.0%	0	.0%	158	100.0%

6. I am: * 10. The country where I have most recently been involved with EAP testing and assessment is: Crosstabulation

			10. The country where I have most recently been involved with EAP testing and assessment is:		
			Rest of The World	United Kingdom	Total
6. I am:	Female	Count	41	62	103
		Expected Count	35.9	67.1	103.0
	Male	Count	14	41	55
		Expected Count	19.1	35.9	55.0
Total		Count	55	103	158
		Expected Count	55.0	103.0	158.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3.254[a]	1	.071		
Continuity Correction[b]	2.652	1	.103		
Likelihood Ratio	3.342	1	.068		
Fisher's Exact Test				.081	.050
N of Valid Cases	158				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 19.15.
b. Computed only for a 2x2 table

(C-S2) Respondent age group and country of most recent involvement with EAP assessment

The '65 and over' category holds an expected count of less than five so the data was recoded in order to create a new category '55 and over' which subsumed the two fields '55-64' and '65 and over'.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
7. My age bracket is: * 10. The country where I have most recently been involved with EAP testing and assessment is:	158	100.0%	0	.0%	158	100.0%

7. My age bracket is: * 10. The country where I have most recently been involved with EAP testing and assessment is: Crosstabulation

			10. The country where I have most recently been involved with EAP testing and assessment is:		
			Rest of The World	United Kingdom	Total
7. My age bracket is:	22 to 34	Count	6	14	20
		Expected Count	7.0	13.0	20.0
	35 to 44	Count	13	34	47
		Expected Count	16.4	30.6	47.0
	45 to 54	Count	19	31	50
		Expected Count	17.4	32.6	50.0
	55 and over	Count	17	24	41
		Expected Count	14.3	26.7	41.0
Total		Count	55	103	158
		Expected Count	55.0	103.0	158.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2.287[a]	3	.515
Likelihood Ratio	2.308	3	.511
N of Valid Cases	158		

0 cells (.0%) have expected count less than 5. The minimum expected count is 6.96.

(C-S3) Respondent institution type of most recent involvement with EAP assessment and country of most recent involvement with EAP assessment

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
11. The most recent institution where I have been involved with EAP testing and assessment can best be described as a(n): * 10. The country where I have most recently been involved with EAP testing and assessment is:	158	100.0%	0	.0%	158	100.0%

11. The most recent institution where I have been involved with EAP testing and assessment, can best be described as a(n): * 10. The country where I have most recently been involved with EAP testing and assessment is: Crosstabulation

			10. The country where I have most recently been involved with EAP testing and assessment is:		
			Rest of The World	United Kingdom	Total
11. The most recent institution where I have been involved with EAP testing and assessment, can best be described as a(n):	University	Count	51	85	136
		Expected Count	47.3	88.7	136.0
	University partnership institution (eg INTO, Kaplan, Navitas)	Count	4	18	22
		Expected Count	7.7	14.3	22.0
Total		Count	55	103	158
		Expected Count	55.0	103.0	158.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3.114[a]	1	.078		
Continuity Correction[b]	2.321	1	.128		
Likelihood Ratio	3.412	1	.065		
Fisher's Exact Test				.094	.060
N of Valid Cases	158				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 7.66.
b. Computed only for a 2x2 table

(C-S4) Respondent time elapsed since starting EAP practice and country of most recent involvement with EAP assessment

The ‘0-23 months ago’ category holds an expected count of less than five so the data was recoded in order to create a new category ‘0-5 years’ which subsumed the two fields ‘0-23 months ago’ and ‘2-5 years ago’.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
12. I first worked in the field of EAP: * 10. The country where I have most recently been involved with EAP testing and assessment is:	158	100.0%	0	.0%	158	100.0%

12. I first worked in the field of EAP: * 10. The country where I have most recently been involved with EAP testing and assessment is: Crosstabulation

			10. The country where I have most recently been involved with EAP testing and assessment is:		
			Rest of The World	United Kingdom	Total
12. I first worked in the field of EAP:	0-5 years ago	Count	18	24	42
		Expected Count	14.6	27.4	42.0
	Over 10 years ago	Count	19	37	56
		Expected Count	19.5	36.5	56.0
	Over 20 years ago	Count	10	18	28
		Expected Count	9.7	18.3	28.0
	Over 5 years ago	Count	8	24	32
		Expected Count	11.1	20.9	32.0
Total		Count	55	103	158
		Expected Count	55.0	103.0	158.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2.585[a]	3	.460
Likelihood Ratio	2.624	3	.453
N of Valid Cases	158		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 9.75.

(C-S5) Respondent gender and institution type of most recent involvement with EAP assessment

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
6. I am: * 11. The most recent institution where I have been involved with EAP testing and assessment, can best be described as a(n):	158	100.0%	0	.0%	158	100.0%

6. I am: * 11. The most recent institution where I have been involved with EAP testing and assessment, can best be described as a(n): Crosstabulation

			11. The most recent institution where I have been involved with EAP testing and assessment, can best be described as a(n):		
			University	University partnership institution (eg INTO, Kaplan, Navitas)	Total
6. I am:	Female	Count	90	13	103
		Expected Count	88.7	14.3	103.0
	Male	Count	46	9	55
		Expected Count	47.3	7.7	55.0
Total		Count	136	22	158
		Expected Count	136.0	22.0	158.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.419[a]	1	.517		
Continuity Correction[b]	.165	1	.685		
Likelihood Ratio	.411	1	.522		
Fisher's Exact Test				.630	.337
N of Valid Cases	158				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 7.66.
b. Computed only for a 2x2 table

(C-S6) Respondent age group and institution type

Given that three expected counts showed a figure of less than 5 the data was recoded so that both the ‘22 to 34’ and ‘35 to 44’ categories were conflated to create a ’22 to 44’ category and the ‘55 to 64’ and ’65 and over’ categories were combined to create a ’55 and over’ band.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
7. My age bracket is: * 11. The most recent institution where I have been involved with EAP testing and assessment, can best be described as a(n):	158	100.0%	0	.0%	158	100.0%

7. My age bracket is: * 11. The most recent institution where I have been involved with EAP testing and assessment, can best be described as a(n): Crosstabulation

			11. The most recent institution where I have been involved with EAP testing and assessment, can best be described as a(n):		
			University	University partnership institution (eg INTO, Kaplan, Navitas)	Total
7. My age bracket is:	22 to 44	Count	52	15	67
		Expected Count	57.7	9.3	67.0
	45 to 54	Count	46	4	50
		Expected Count	43.0	7.0	50.0
	55 and Over	Count	38	3	41
		Expected Count	35.3	5.7	41.0
Total		Count	136	22	158
		Expected Count	136.0	22.0	158.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6.962[a]	2	.031
Likelihood Ratio	6.932	2	.031
N of Valid Cases	158		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 5.71.

(C-S7) Respondent time elapsed since starting EAP practice and type of institution

Due to the fact that 4 of the expected values in the data set fell below 5, the data was recoded so that the '0-23 months ago' category and the '2-5 years ago' grouping were combined to form '0-5 years'. The remaining three bands were then combined to form a category labelled as 'over 6 years'.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
12. I first worked in the field of EAP: * 11. The most recent institution where I have been involved with EAP testing and assessment, can best be described as a(n):	158	100.0%	0	.0%	158	100.0%

12. I first worked in the field of EAP: * 11. The most recent institution where I have been involved with EAP testing and assessment, can best be described as a(n): Crosstabulation

			11. The most recent institution where I have been involved with EAP testing and assessment, can best be described as a(n):		
			University	University partnership institution (eg INTO, Kaplan, Navitas)	Total
12. I first worked in the field of EAP:	0-5 years ago	Count	34	8	42
		Expected Count	36.2	5.8	42.0
	Over 5 years ago	Count	102	14	116
		Expected Count	99.8	16.2	116.0
Total		Count	136	22	158
		Expected Count	136.0	22.0	158.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.253[a]	1	.263		
Continuity Correction[b]	.738	1	.390		
Likelihood Ratio	1.187	1	.276		
Fisher's Exact Test				.300	.193
N of Valid Cases	158				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 5.85.
b. Computed only for a 2x2 table

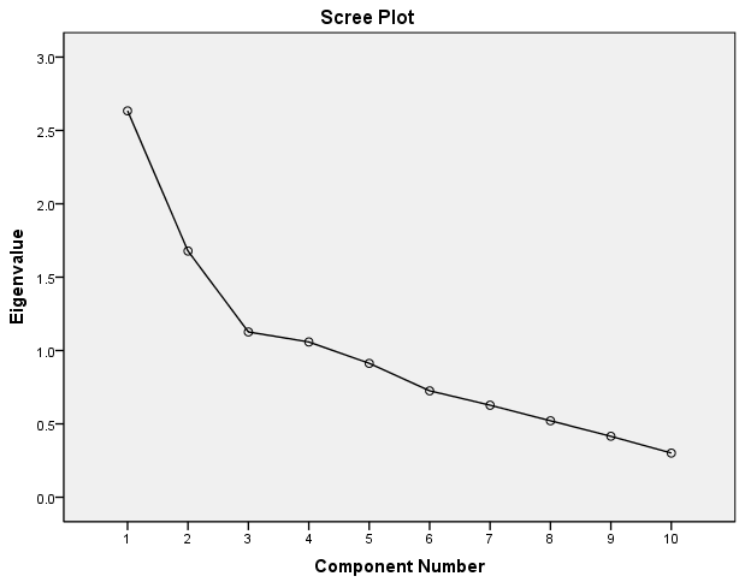
8.7 Factor Analysis & Cronbach's Alpha results for DVAS items AQ13, AQ14, AQ16, AQ19, AQ20, AQ22, AQ23, AQ25 and AQ26 for Population B

8.7.1 Factor Analysis for AQ13A-J

Rotated Component Matrix[a]

	Component			
	1	2	3	4
Q13_a	.070	-.043	-.027	.895
Q13_b	.125	.055	.728	.145
Q13_c	.159	-.164	.741	-.188
Q13_d	.624	-.034	.288	.226
Q13_e	.645	.018	.366	.451
Q13_f	.790	-.054	.046	-.201
Q13_g	.699	-.075	.016	.098
Q13_h	-.420	.487	.409	.169
Q13_i	-.050	.866	.026	-.092
Q13_j	-.015	.800	-.140	.014

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 6 iterations.



8.7.2 Cronbach's Alpha Reliability Statistics for AQ13

Reliability Statistics For AQ13 Factor 1

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.720	.722	4

Item-Total Statistics For Q13 Factor 1

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q13_d	9.05	5.310	.500	.348	.665
Q13_e	9.27	5.027	.637	.437	.576
Q13_f	9.13	6.447	.455	.215	.691
Q13_g	9.13	5.606	.460	.265	.689

Reliability Statistics For AQ13 Factor 2

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.614	.613	3

Item-Total Statistics For AQ13 Factor 2

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q13_h	5.98	2.885	.328	.157	.646
Q13_i	5.59	2.256	.571	.331	.281
Q13_j	5.42	2.818	.384	.229	.567

Reliability Statistics For AQ13 Factor 3

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.420	.446	2

Item-Total Statistics For AQ13 Factor 3

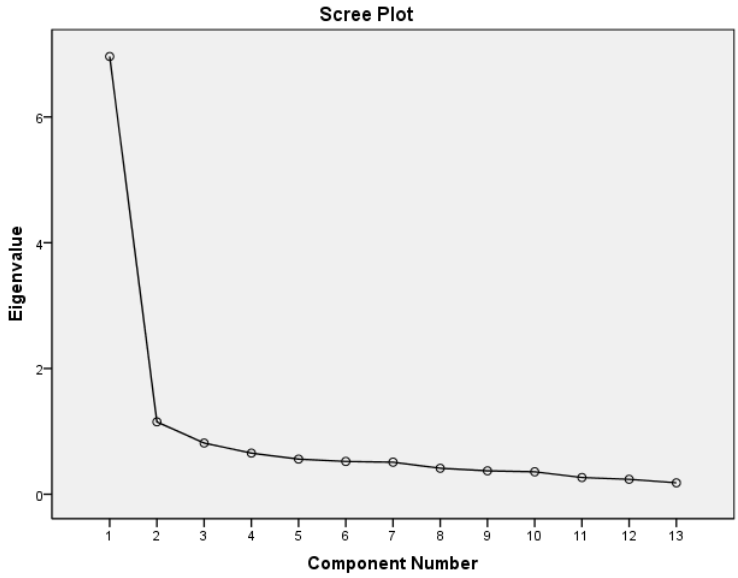
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q13_b	3.13	1.377	.287	.082	.
Q13_c	1.92	.624	.287	.082	.

8.7.3 Factor Analysis for AQ14A-M

Rotated Component Matrix[a]

	Component	
	1	2
Q14_a	.534	.548
Q14_b	.550	.531
Q14_c	.631	.382
Q14_d	.514	.508
Q14_e	.638	.405
Q14_f	.426	.600
Q14_g	.181	.868
Q14_h	.158	.839
Q14_i	.689	.373
Q14_j	.780	.275
Q14_k	.764	.186
Q14_l	.819	.179
Q14_m	.769	.198

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations.



8.7.4 Cronbach's Alpha Reliability Statistics for AQ14

After Factor Analysis and researcher intuition after consideration of the data 2 factors were retained with AQ14a and AQ14f being subsumed within Factor1.

Reliability Statistics For Factor 1

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.922	.923	11

Item-Total Statistics For AQ14 Factor 1

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q14_a	25.80	33.891	.695	.610	.914
Q14_b	25.99	34.000	.689	.537	.915
Q14_c	26.07	34.167	.680	.514	.915
Q14_d	25.66	33.358	.647	.442	.917
Q14_e	25.93	33.632	.711	.569	.914
Q14_f	25.74	33.263	.618	.405	.919
Q14_i	25.61	32.316	.716	.631	.913
Q14_j	25.94	32.360	.748	.704	.912
Q14_k	26.46	33.982	.680	.527	.915
Q14_l	25.89	32.797	.727	.607	.913
Q14_m	25.97	33.967	.694	.547	.915

Reliability Statistics For Factor 2

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.816	.817	2

Item-Total Statistics For Factor 2

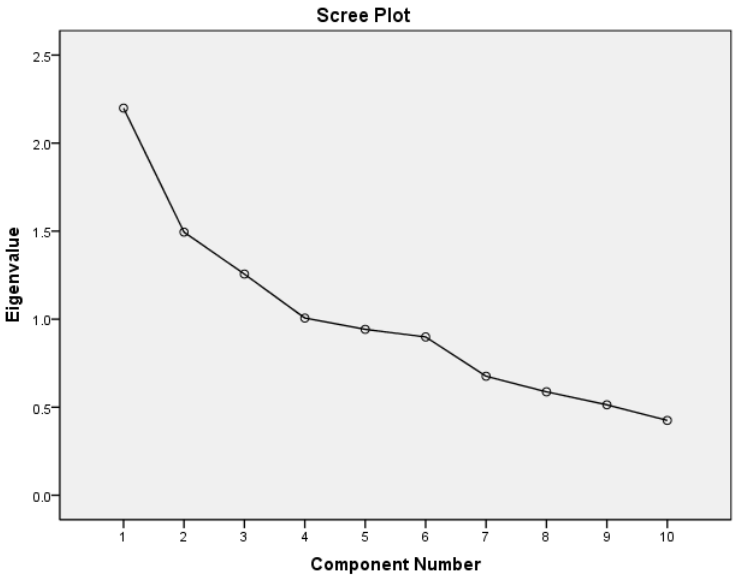
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q14_g	4.16	1.000	.691	.478	.
Q14_h	3.34	1.154	.691	.478	.

8.7.5 Factor Analysis for AQ16A-J

Rotated Component Matrix[a]

	Component			
	1	2	3	4
Q16_a	.649	-.381	.041	.403
Q16_b	.292	.529	-.236	-.011
Q16_c	.033	-.478	.082	.013
Q16_d	-.115	-.002	-.092	.856
Q16_e	-.113	.559	.186	.641
Q16_f	.735	-.053	.171	-.115
Q16_g	.670	.181	.199	-.260
Q16_h	-.096	.677	.017	.091
Q16_i	.133	.017	.835	.029
Q16_j	-.196	.232	-.761	.062

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 7 iterations.



8.7.6 Cronbach's Alpha Reliability Statistics for AQ16

Reliability Statistics AQ16 Factor 1

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.538	.536	3

Item-Total Statistics AQ16 Factor 1

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q16_a	4.49	2.124	.313	.105	.493
Q16_f	5.38	1.613	.420	.178	.312
Q16_g	5.68	2.284	.329	.121	.472

Reliability Statistics AQ16 for Factor 2

Cronbach's Alpha[a]	Cronbach's Alpha Based on Standardized Items[a]	N of Items
-.081	-.065	3

a. The value is negative due to a negative average covariance among items. This violates reliability model assumptions. You may want to check item codings.

Item-Total Statistics AQ16 for Factor 2

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q16_b	4.53	1.499	.000	.017	-.147[a]
Q16_c	4.56	1.408	-.107	.012	.178
Q16_h	4.47	1.270	.008	.013	-.198[a]

a. The value is negative due to a negative average covariance among items. This violates reliability model assumptions. You may want to check item codings.

Reliability Statistics For AQ16 Factor 3

Cronbach's Alpha[a]	Cronbach's Alpha Based on Standardized Items[a]	N of Items
-1.507	-1.513	2

a. The value is negative due to a negative average covariance among items. This violates reliability model assumptions. You may want to check item codings.

Item-Total Statistics For AQ16 Factor 3

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q16_i	2.28	.597	-.431	.186	.
Q16_j	3.94	.684	-.431	.186	.

Reliability Statistics AQ16 Factor 4

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.566	.567	2

Item-Total Statistics For AQ16 Factor 4

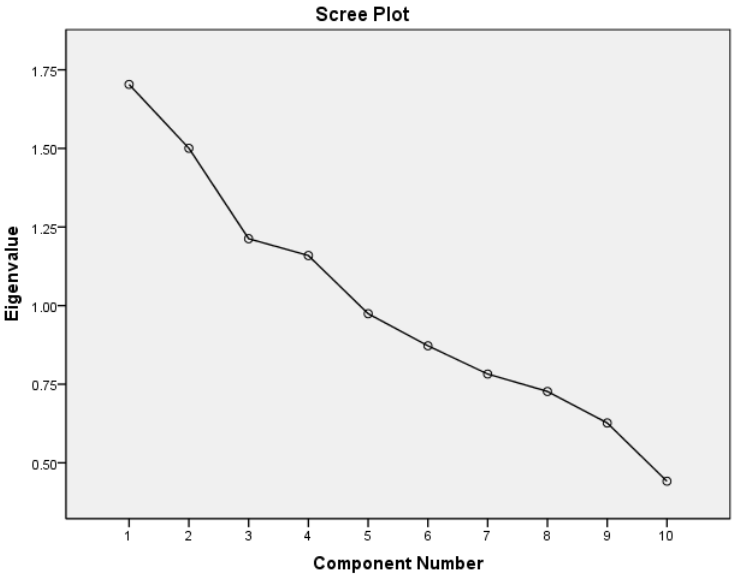
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q16_d	2.88	.884	.395	.156	.
Q16_e	2.46	1.014	.395	.156	.

8.7.7 Factor Analysis for AQ19A-J

Rotated Component Matrix[a]

	Component			
	1	2	3	4
Q19_a	.442	.064	-.196	.537
Q19_b	-.193	-.676	.017	.221
Q19_c	-.183	.687	-.113	.160
Q19_d	-.064	.533	.208	.097
Q19_e	.831	-.106	.096	-.012
Q19_f	.814	.017	.013	.043
Q19_g	.096	-.086	.796	-.046
Q19_h	-.044	.111	.594	.439
Q19_i	-.028	.002	.059	.776
Q19_j	-.007	.340	.474	-.297

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 6 iterations.



8.7.8 Cronbach's Alpha Reliability Statistics for AQ19

Reliability Statistics For Factor1 AQ19

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.642	.642	2

Item-Total Statistics For Factor 1 AQ19

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q19_e	2.41	.854	.473	.224	.
Q19_f	3.07	.779	.473	.224	.

Reliability Statistics For Factor 2 AQ19

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.339	.340	3

Item-Total Statistics For Factor 2 AQ19

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q19_b	5.36	2.359	.203	.042	.234
Q19_c	4.72	1.973	.208	.044	.225
Q19_d	5.37	2.643	.172	.030	.295

Reliability Statistics For Factor 3 AQ19

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.084	.051	3

Item-Total Statistics For Factor 3AQ19

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q19_g	5.62	1.817	.191	.045	-.337[a]
Q19_h	4.99	1.529	.060	.069	.000
Q19_i	5.51	2.876	-.114	.026	.338

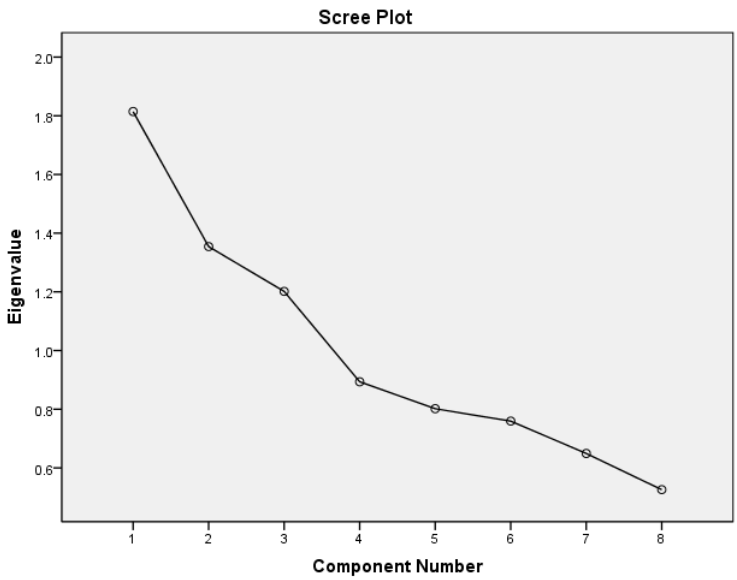
a. The value is negative due to a negative average covariance among items. This violates reliability model assumptions. You may want to check item codings.

8.7.9 Factor Analysis for AQ20A-H

Rotated Component Matrix[a]

	Component		
	1	2	3
Q20_a	-.108	.838	.032
Q20_b	.227	.022	-.700
Q20_c	.084	.097	.672
Q20_d	.255	-.155	.617
Q20_e	-.051	.839	-.074
Q20_f	.503	-.215	-.064
Q20_g	.777	-.056	-.011
Q20_h	-.715	-.066	-.190

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 4 iterations.



8.7.10 Cronbach's Alpha Reliability Statistics for AQ20

Reliability Statistics AQ20 Factor 1

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.461	.456	3

Item-Total Statistics AQ20 Factor 1

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q20_f	8.16	1.718	.208	.051	.484
Q20_g	7.94	1.295	.365	.137	.205
Q20_h	7.56	1.433	.285	.104	.361

Reliability Statistics For AQ20 Factor 2

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.623	.634	2

Item-Total Statistics For AQ20 Factor 2

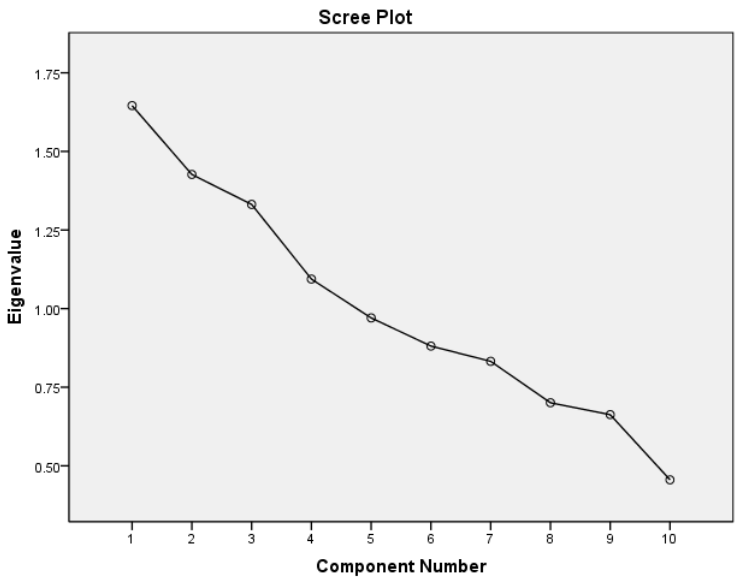
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q20_a	1.09	.112	.465	.216	.
Q20_e	1.18	.176	.465	.216	.

8.7.11 Factor Analysis for AQ22A-J

Rotated Component Matrix[a]

	Component			
	1	2	3	4
Q22_a	.174	.695	.201	-.058
Q22_b	.071	.597	-.347	-.075
Q22_c	-.388	.581	.164	.047
Q22_d	.819	-.113	.217	.021
Q22_e	.744	.179	-.043	.226
Q22_f	.099	.138	-.086	.804
Q22_g	.165	-.254	-.154	.538
Q22_h	.017	.054	.678	-.141
Q22_i	.130	.068	.615	-.064
Q22_j	.342	.178	-.486	-.458

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 5 iterations.



8.7.12 Cronbach's Alpha Reliability Statistics for AQ22

Reliability Statistics For AQ22 Factor 1

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.588	.590	2

Item-Total Statistics For AQ22 Factor 1

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q22_d	1.91	.341	.419	.175	.
Q22_e	2.01	.427	.419	.175	.

Reliability Statistics For AQ22 Factor 2

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.296	.296	3

Item-Total Statistics for AQ22 Factor 2

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q22_a	5.39	1.628	.202	.041	.135
Q22_b	5.39	1.769	.154	.028	.240
Q22_c	5.22	1.724	.138	.021	.276

Reliability Statistics for AQ22 Factor 3

Cronbach's Alpha[a]	Cronbach's Alpha Based on Standardized Items[a]	N of Items
-.052	-.022	3

a. The value is negative due to a negative average covariance among items. This violates reliability model assumptions. You may want to check item codings.

Item-Total Statistics for AQ22 Factor 3

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q22_h	5.25	1.818	.048	.030	-.217[a]
Q22_i	5.39	1.654	.022	.033	-.175[a]
Q22_j	5.17	1.837	-.119	.014	.273

a. The value is negative due to a negative average covariance among items. This violates reliability model assumptions. You may want to check item codings.

Reliability Statistics for AQ22 Factor 4

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.290	.294	2

Item-Total Statistics for AQ22 Factor 4

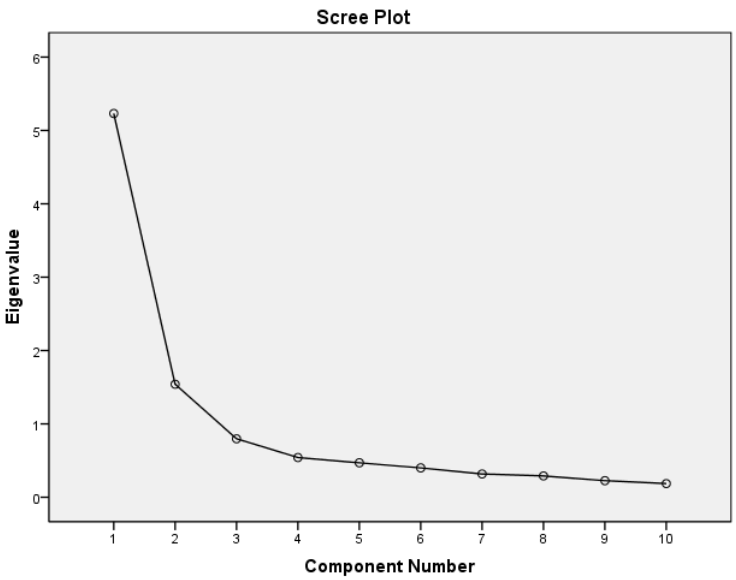
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q22_f	2.37	.973	.173	.030	.
Q22_g	3.29	1.443	.173	.030	.

8.7.13 Factor Analysis for AQ23A-J

Rotated Component Matrix[a]

	Component	
	1	2
Q23_a	.667	.355
Q23_b	.712	.413
Q23_c	.194	.788
Q23_d	.861	.048
Q23_e	.180	.861
Q23_f	.881	.123
Q23_g	.127	.833
Q23_h	.685	.245
Q23_i	.359	.752
Q23_j	.569	.544

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations.



8.7.14 Cronbach's Alpha Reliability Statistics for AQ23

Reliability Statistics AQ23 Factor 1

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.863	.872	6

Item-Total Statistics AQ23 Factor1

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q23_a	10.59	20.701	.643	.551	.842
Q23_b	10.71	21.055	.720	.635	.833
Q23_d	10.94	21.428	.674	.550	.840
Q23_f	10.71	19.515	.757	.638	.822
Q23_h	10.23	19.318	.602	.426	.854
Q23_j	9.83	18.991	.623	.414	.850

Reliability Statistics AQ23 Factor 2

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.861	.865	4

Item-Total Statistics AQ23 Factor 2

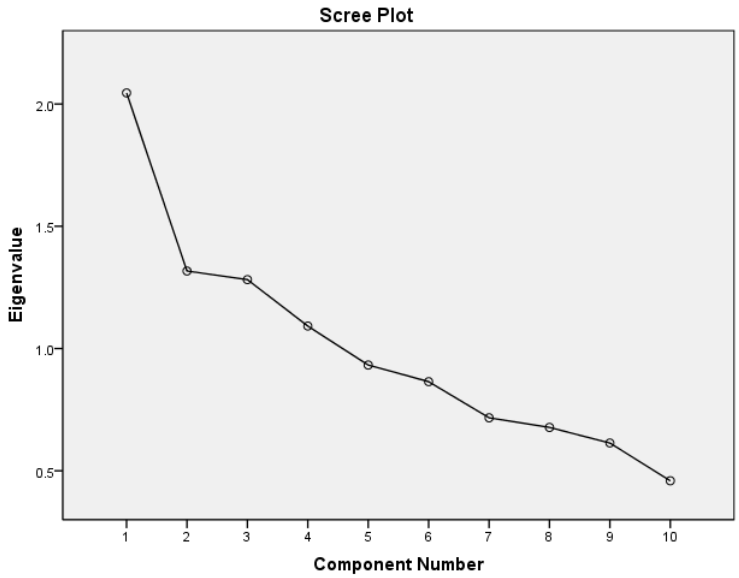
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q23_c	10.72	11.732	.674	.457	.837
Q23_g	9.84	12.159	.712	.530	.825
Q23_e	9.95	11.095	.773	.608	.797
Q23_i	10.57	10.626	.691	.494	.835

8.7.15 Factor Analysis for AQ25A-J

Rotated Component Matrix[a]

	Component			
	1	2	3	4
Q25_a	.753	.124	-.081	.168
Q25_b	.555	-.181	-.018	.337
Q25_c	-.377	.553	-.009	-.188
Q25_d	-.250	-.150	-.491	.486
Q25_e	.649	-.153	.144	-.091
Q25_f	.696	.055	-.121	-.303
Q25_g	.069	.827	.028	.069
Q25_h	-.002	.332	.685	.300
Q25_i	.060	.041	.047	.721
Q25_j	-.155	-.361	.698	-.131

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 7 iterations.



8.7.16 Cronbach's Alpha Reliability Statistics for AQ25

Reliability Statistics AQ25 Factor 1

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.617	.622	4

Item-Total Statistics AQ25 Factor 1

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q25_a	10.93	4.334	.488	.263	.485
Q25_b	11.06	4.501	.342	.177	.588
Q25_e	10.82	4.447	.392	.161	.551
Q25_f	10.93	4.280	.377	.209	.563

Reliability Statistics For AQ25 Factor 2

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.475	.476	2

Item-Total Statistics For AQ25 factor 2

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q25_c	2.99	.764	.312	.097	.
Q25_g	3.39	.863	.312	.097	.

Reliability Statistics For AQ25 Factor3

Cronbach's Alpha[a]	Cronbach's Alpha Based on Standardized Items[a]	N of Items
-.134	-.127	3

a. The value is negative due to a negative average covariance among items. This violates reliability model assumptions. You may want to check item codings.

Item-Total Statistics For AQ25 Factor 3

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q25_d	5.97	2.139	-.152	.026	.203
Q25_i	6.25	1.643	.009	.018	-.330[a]
Q25_j	6.53	1.983	-.021	.031	-.194[a]

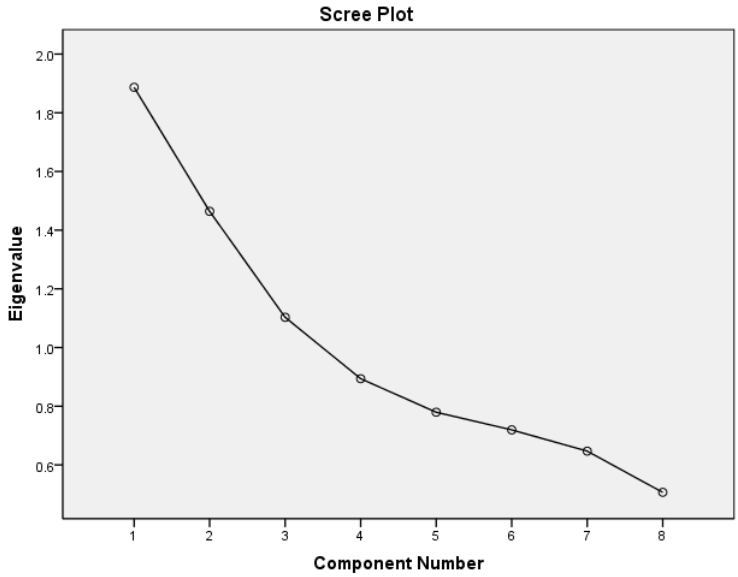
a. The value is negative due to a negative average covariance among items. This violates reliability model assumptions. You may want to check item codings.

8.7.17 Factor Analysis for AQ26A-H

Rotated Component Matrix[a]

	Component		
	1	2	3
Q26_a	-.103	.479	.409
Q26_b	-.010	.710	-.106
Q26_c	.027	-.026	-.821
Q26_d	.057	.779	.085
Q26_e	.689	.442	-.007
Q26_f	.666	.012	.238
Q26_g	.707	-.186	-.184
Q26_h	.452	-.059	.539

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 5 iterations.



8.7.18 Cronbach's Alpha for AQ26A-H

Reliability Statistics For AQ26 Factor 1

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.537	.538	4

Item-Total Statistics For AQ26 Factor 1

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q26_e	7.60	2.047	.401	.164	.394
Q26_f	7.22	1.978	.363	.149	.430
Q26_g	5.65	2.269	.270	.081	.513
Q26_h	7.00	2.649	.277	.081	.506

Reliability Statistics For AQ26 Factor 2

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.478	.471	3

Item-Total Statistics For AQ26 Factor 2

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q26_a	3.53	1.909	.222	.055	.494
Q26_b	3.30	1.501	.312	.113	.354
Q26_d	3.21	1.180	.377	.144	.225

8.7.19 Cronbach's Alpha for factors combined from AQ14 Factor 2 and AQ23Factors 1 and 2

AQ23 Factor 1 with the addition of item G from AQ14 Factor 2

Reliability Statistics For AQ14 Factor 1 + Item AQ23g

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.870	.876	7

Item-Total Statistics For AQ14 Factor 1 + Item AQ23g

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q23_a	13.90	27.108	.651	.559	.851
Q23_b	14.03	27.555	.721	.636	.844
Q23_d	14.25	27.968	.677	.562	.849
Q23_f	14.02	25.993	.738	.641	.839
Q23_h	13.55	25.465	.618	.453	.858
Q23_j	13.14	25.156	.633	.427	.856
Q14_g	12.60	28.267	.552	.326	.863

AQ23 Factor 2 with the addition of item H from AQ14 Factor 2

Reliability Statistics For AQ14 factor 2 + Item AQ23h

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.868	.871	5

Item-Total Statistics For AQ14 Factor 2 + Item AQ23h

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q23_c	14.88	16.692	.691	.481	.840
Q23_g	14.01	17.353	.707	.530	.837
Q23_e	14.11	15.864	.795	.645	.813
Q23_i	14.73	15.556	.689	.495	.845
Q14_h	13.69	19.377	.606	.394	.861

8.7.20 Cronbach's Alpha for factors combined from AQ16 Factor 4 & AQ19 Factors 1

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.632	.631	4

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q19_e	7.75	4.865	.274	.244	.652
Q19_f	8.41	3.721	.599	.393	.421
Q16_d	8.35	4.077	.398	.255	.574
Q16_e	7.94	4.289	.395	.199	.574

8.8 Interview Interactional Profiles IA-IY

Interviewee Code	I_A	Interview Mode	Skype	Institution Type	University
Country of Practice	UK	Gender	F	Age group	35-44
Length of Interview	19 Minutes 23 Seconds			Years of practice	5 years +
Features of Interaction	Interviewee/Interviewer Status	Peer-to-Peer Respondent seemed aware of interviewer having higher level of experience.			
	Interviewee Engagement	Engaged with subject but notably inexperienced in some areas			
	Other Comments	High level of enthusiasm but does not consider self as an expert. Time pressured interview			

Interviewee Code	I_B	Interview Mode	Face-to-Face	Institution Type	University
Country of Practice	UK	Gender	F	Age group	45-54
Length of Interview	5 Minutes 42 Seconds			Years of practice	10 years+
Features of Interaction	Interviewee/Interviewer Status	Peer-to-Peer			
	Interviewee Engagement	Interviewee showed only a moderate level of engagement with the topic. A number of short answers provided			
	Other Comments	Time pressured interview			

Interviewee Code	I_C	Interview Mode	Face-to-Face	Institution Type	University
Country of Practice	UK	Gender	M	Age group	45-54
Length of Interview	7 Minutes 3 Seconds			Years of practice	10 years+
Features of Interaction	Interviewee/Interviewer Status	Peer-to-Peer			
	Interviewee Engagement	Engaged with subject			
	Other Comments	Time pressured interview Interviewee appeared humble and positioned themselves as less than expert in a number of areas			

Interviewee Code	I_D	Interview Mode	Face-to-Face	Institution Type	University
Country of Practice	UK	Gender	M	Age group	22-34
Length of Interview	11 Minutes 34 Seconds			Years of practice	2-5 years ago
Features of Interaction	Interviewee/Interviewer Status	Peer-to-Peer Although interviewee is a junior colleague at the same institution as the interviewer the interview tone and delivery maintained a peer-to- peer tone.			
	Interviewee Engagement	Engaged with the subject area			
	Other Comments	Seemed confident and aware of key issues and potential effects on stakeholders.			

Interviewee Code	I_E	Interview Mode	Face-to-face	Institution Type	University
Country of Practice	UK	Gender	F	Age group	55-64
Length of Interview	12 Minutes 10 Seconds			Years of practice	20 years+
Features of Interaction	Interviewee/Interviewer Status	Peer-to-Peer Former colleague at a different institution.			
	Interviewee Engagement	Neither engaged nor disengaged with subject area.			
	Other Comments	Experienced but did not present self as an expert rather reliant on other colleagues with expertise.			

Interviewee Code	I_F	Interview Mode	Face-to-Face	Institution Type	Partnership
Country of Practice	UK	Gender	F	Age group	35-44
Length of Interview	9 Minutes 52 Seconds			Years of practice	2-5 years ago
Features of Interaction	Interviewee/Interviewer Status	Peer-to-Peer Although interviewee projected relative inexperience.			
	Interviewee Engagement	Neither engaged nor disengaged with topic			
	Other Comments	Interviewee seemed to lack confidence with the topic and seemed to feel marginalised by their institution.			

Interviewee Code	I_G	Interview Mode	Face-to-Face	Institution Type	University
Country of Practice	UK	Gender	F	Age group	55-64
Length of Interview	17 Minutes 17 Seconds			Years of practice	10 years+
Features of Interaction	Interviewee/Interviewer Status	Peer-to-Peer / Junior-to-Senior Tone of interaction indicated interviewees wish to appear senior			
	Interviewee Engagement	Neither engaged nor disengaged			
	Other Comments	Interviewee very confident of their own approach despite flaws which were self-acknowledged.			

Interviewee Code	I_H	Interview Mode	Skype	Institution Type	University
Country of Practice	US	Gender	F	Age group	22-34
Length of Interview	13 minutes 57 Seconds			Years of practice	2-5years
Features of Interaction	Interviewee/Interviewer Status	Peer-to-Peer/Junior-to-senior Despite interviewer efforts the interviewees lack of confidence with the topic was notable			
	Interviewee Engagement	Engaged from the perspective of student needs			
	Other Comments	Interviewee did not appear confident talking about the subject and I some situations did not fully understand the question.			

Interviewee Code	I_I	Interview Mode	Face-to-Face	Institution Type	University
Country of Practice	UK	Gender	F	Age group	65+
Length of Interview	8 Minutes 39 Seconds			Years of practice	20 years+
Features of Interaction	Interviewee/Interviewer Status	Peer-to-Peer			
	Interviewee Engagement	Engaged with the topic due to concern for quality			
	Other Comments	Does not present themselves as an expert in the field- lacks confidence in own ability			

Interviewee Code	I_J	Interview Mode	Face-to-Face	Institution Type	University
Country of Practice	China	Gender	M	Age group	35-44
Length of Interview	15 Minutes 50 Seconds			Years of practice	5 years+
Features of Interaction	Interviewee/Interviewer Status	Peer-to-Peer Although the interviewee did not appear wholly relaxed.			
	Interviewee Engagement	Engaged with subject			
	Other Comments	Seemed keen to demonstrate expertise			

Interviewee Code	I_K	Interview Mode	Skype	Institution Type	University
Country of Practice	UK	Gender	M	Age group	22-34
Length of Interview	25 Minutes 20 Seconds			Years of practice	2-5 years
Features of Interaction	Interviewee/Interviewer Status	Peer-to-Peer			
	Interviewee Engagement	Very engaged and keen to contribute from a range of experience			
	Other Comments	Noted a number of instances of internal politics in institutions where they worked			

Interviewee Code	I_L	Interview Mode	Skype	Institution Type	Partnership
Country of Practice	UK	Gender	F	Age group	54-64
Length of Interview	16 Minutes 36 Seconds			Years of practice	20 years+
Features of Interaction	Interviewee/Interviewer Status	Junior-to-Senior Interviewee expertise established a tone which positioned them as senior to the interviewer with regard to expertise in the field			
	Interviewee Engagement	Very engaged with topic			
	Other Comments	Seemed to demonstrate a high level of expertise in the field			

Interviewee Code	I_M	Interview Mode	Skype	Institution Type	University
Country of Practice	UK	Gender	F	Age group	22-34
Length of Interview	22 Minutes 36 Seconds			Years of practice	0-23 months
Features of Interaction	Interviewee/Interviewer Status	Peer-to-peer			
	Interviewee Engagement	Engaged but did not position themselves as an expert			
	Other Comments	Interviewee recognised constraints of real practice in action			

Interviewee Code	I_N	Interview Mode	Face-to-Face	Institution Type	Partnership
Country of Practice	China	Gender	F	Age group	35-44
Length of Interview	47 Minutes 4 Seconds			Years of practice	10 years+
Features of Interaction	Interviewee/Interviewer Status	Peer-to-Peer			
	Interviewee Engagement	Very engaged			
	Other Comments	Very fluent speaker, difficult to move from question to question, speaker appeared disillusion with current institution			

Interviewee Code	I_O	Interview Mode	Skype	Institution Type	University
Country of Practice	Australia	Gender	F	Age group	45-54
Length of Interview	13 Minutes 1 Second			Years of practice	10 years+
Features of Interaction	Interviewee/Interviewer Status	Peer-to-Peer The start of the interview the interviewee seemed to lack confidence and grew with confidence as interview progressed			
	Interviewee Engagement	Neither engaged nor disengaged with topic			
	Other Comments	Referred to examples in IELTS contexts			

Interviewee Code	I_P	Interview Mode	Skype	Institution Type	Partnership
Country of Practice	UK	Gender	F	Age group	35-44
Length of Interview	13 Minutes 36 Seconds			Years of practice	5 years +
Features of Interaction	Interviewee/Interviewer Status	Peer-to-Peer			
	Interviewee Engagement	Engaged			
	Other Comments	Referred to high stakes nature of testing. Referred to relative low skills in some areas			

Interviewee Code	I_Q	Interview Mode	Skype	Institution Type	University
Country of Practice	UK	Gender	F	Age group	55-64
Length of Interview	20 Minutes 56 Seconds			Years of practice	20 years+
Features of Interaction	Interviewee/Interviewer Status	Peer-to-peer			
	Interviewee Engagement	Neither engaged nor disengaged			
	Other Comments	Referred to lack of skills in some respects due to current role in institution			

Interviewee Code	I_R	Interview Mode	Face-to-Face	Institution Type	University
Country of Practice	UK	Gender	F	Age group	35-44
Length of Interview	15 Minutes 56 Seconds			Years of practice	10 years+
Features of Interaction	Interviewee/Interviewer Status	Peer-to-peer			
	Interviewee Engagement	Neither engaged nor disengaged. Recognises importance of area but it is not a particular area of key interest			
	Other Comments	Has recently developed an interest through own study			

Interviewee Code	I_S	Interview Mode	Face-to-Face	Institution Type	University
Country of Practice	UK	Gender	F	Age group	45-54
Length of Interview	23 Minutes 43 Seconds			Years of practice	10 years+
Features of Interaction	Interviewee/Interviewer Status	Peer-to-Peer			
	Interviewee Engagement	Engaged, Very fluent speaker on topic from a practitioner perspective rather than expert			
	Other Comments	Focused on student needs. Refers to development of expertise on the job			

Interviewee Code	I_T	Interview Mode	Face-to-Face	Institution Type	University
Country of Practice	UK	Gender	M	Age group	35-44
Length of Interview	11 Minutes 22 Seconds			Years of practice	10 years+
Features of Interaction	Interviewee/Interviewer Status	Peer-to-peer			
	Interviewee Engagement	Disengaged – seemed very cynical			
	Other Comments	Very confident of own abilities, critical of institutions Some quite short answers			

Interviewee Code	I_U	Interview Mode	Skype	Institution Type	University
Country of Practice	UK	Gender	M	Age group	45-54
Length of Interview	12 minutes 53 Seconds			Years of practice	5 years+
Features of Interaction	Interviewee/Interviewer Status	Peer-to-Peer Despite interviewee's efforts to appear humble			
	Interviewee Engagement	Very engaged			
	Other Comments	Very experienced yet concerned about own ability			

Interviewee Code	I_V	Interview Mode	Skype	Institution Type	University
Country of Practice	UK	Gender	F	Age group	55-64
Length of Interview	30 Minutes 2 Seconds			Years of practice	0-23 Months
Features of Interaction	Interviewee/Interviewer Status	Peer-to-peer			
	Interviewee Engagement	Engaged yet inexperienced in some areas			
	Other Comments	Very fluent speaker. Difficult to introduce new questions. Some questions misunderstood or answered with off-topic responses			

Interviewee Code	I_W	Interview Mode	Skype	Institution Type	University
Country of Practice	UK	Gender	F	Age group	45-54
Length of Interview	18 Minutes 13 Seconds			Years of practice	10 years+
Features of Interaction	Interviewee/Interviewer Status	Senior-to-junior Interviewee showed a high level of expertise which demonstrated their high level of proficiency during the interview			
	Interviewee Engagement	Engaged			
	Other Comments	Demonstrated a high level of expertise in the field			

Interviewee Code	I_X	Interview Mode	Skype	Institution Type	University
Country of Practice	France	Gender	M	Age group	55-64
Length of Interview	13 Minutes 14 Seconds			Years of practice	5 years+
Features of Interaction	Interviewee/Interviewer Status	Peer-to-peer			
	Interviewee Engagement	Moderate engagement			
	Other Comments	Interviewee acknowledged lack of own skills in assessment			

Interviewee Code	I_Y	Interview Mode	Skype	Institution Type	University
Country of Practice	UK	Gender	M	Age group	45-54
Length of Interview	32 Minutes 25 Seconds			Years of practice	10 years+
Features of Interaction	Interviewee/Interviewer Status	Junior-to-senior -despite friendliness of interviewee their level of skill and experience showed them to be senior to interviewer			
	Interviewee Engagement	Very engaged			
	Other Comments	The interviewee appeared very highly skilled and experienced in the field			