BRITISH COUNCIL

Aptis

Forward thinking
English testing

Aptis test development approach

Aptis technical report (ATR-1)

Barry O'Sullivan, British Council

August 2012

Contents

Aptis test development approach 1

1. Introduction 1

2. Theoretical basis of the Aptis test system 1

2.1. Overview of the model 1

2.2. Application of the model 4

2.2.1. The population 4

2.2.2. The test system 4

2.2.3. The scoring system 7

2.3. Test development model 8

2.4. The concept phase 8

2.4.1. The design phase 9

2.4.2. The construction stage 10

2.4.3. The launch phase 10

References 11

Aptis test development approach

1. Introduction

The Aptis test system is designed to offer users an alternative to currently available high-stakes certificated examinations. The quality assurance and test security requirements of high-stakes examinations often make them fixed in terms of content, and quite expensive for the user.

The fact that the Aptis system is not a certificated test means that elements of the quality assurance system normally associated with high-stakes examinations (such as those related to test administration and invigilation) are not the responsibility of Aptis, but of the system user. It is also envisaged that users will be encouraged to work with Aptis to generate evidence in support of the use of the system for the purpose intended by that user. This policy is in line with O'Sullivan and Weir (2011) and Cizek (2011), who argue against conceptualisation of consequence as an aspect of validity. Supporters of Messick see the consequential aspect of validity as the responsibility of the developer, whereas the approach applied in Aptis sees it as the joint responsibility of the test user and the test developer.

In this document, we present an overview of the Aptis development process. The process was based from its inception on an operational model of test validation and also on a related model of development. These models are presented and explored below and the process outlined and discussed.

2. Theoretical basis of the Aptis test system

The theoretical model of test validation which drives the Aptis test system is based on the modifications to the Weir (2005) model suggested by O'Sullivan (2011) and O'Sullivan and Weir (2011). An overview of this model is presented here.

2.1. Overview of the model

As can be seen from Figure 1, and the explanatory table (Table 2), the model of validation proposed by O'Sullivan (2011) and supported by O'Sullivan and Weir (2011) is based around three basic elements. These are the test taker, the test system and the scoring system. The core of the validation argument is therefore the focus on how these three elements combine to result in a measure of candidate performance which is meaningful in terms of the underlying trait or ability being assessed.

Figure 1: A reconceptualisation of Weir's socio-cognitive framework

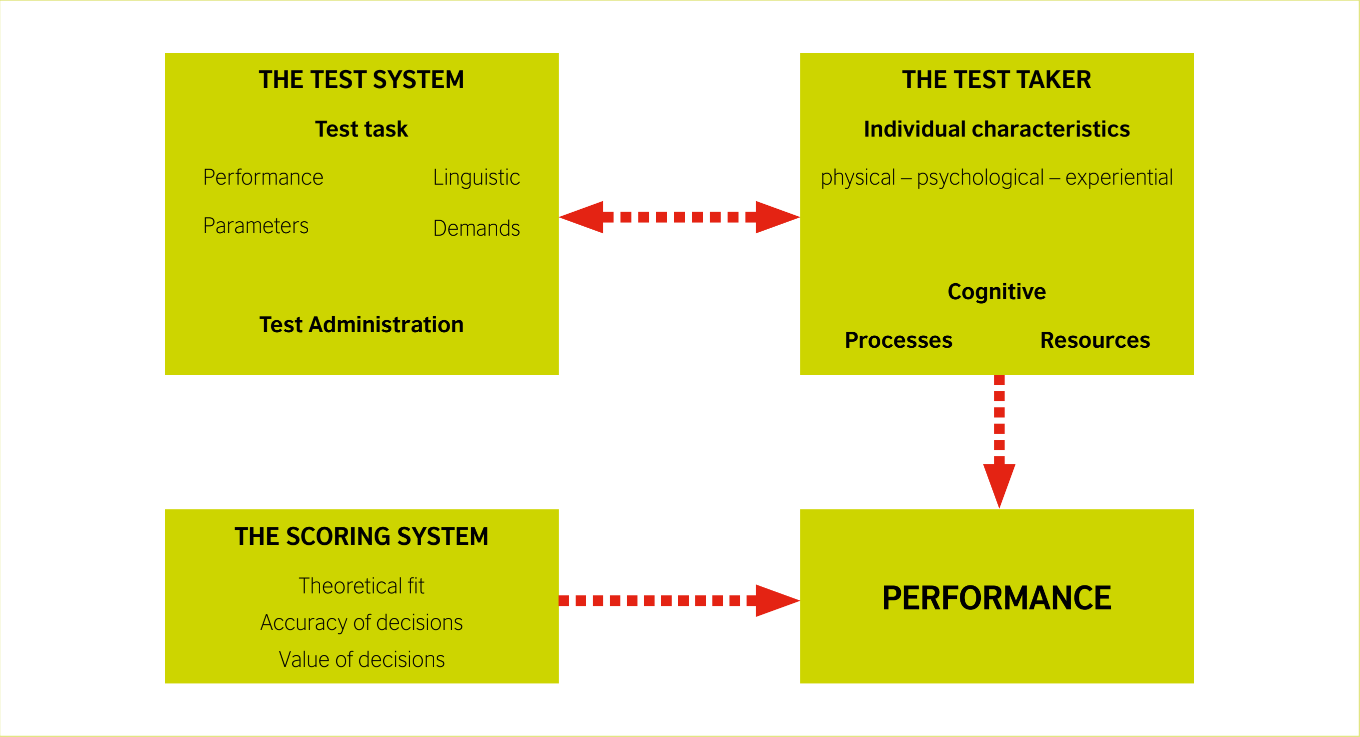


Table 1: Model details (from O'Sullivan, 2011)

The test-taker		
Individual characteristics	Physical	Includes things such as age and gender as well as both short-term ailments (such as cold, toothache) and longer-term disabilities (e.g. dyslexia, limited hearing or sight).
	Psychological	Includes things like memory, personality, cognitive style, affective schemata, concentration, motivation and emotional state.
	Experiential	Includes education (formal and informal) as well as experience of the examination and factors such as target language country residence.
Cognitive	Processes	Cognitive and metacognitive processing.
	Resources	Relates to knowledge of content and to language ability.
The test system		
Test task	Performance parameters	These are parameters such as timing, preparation, score weighting, and knowledge of how performance will be scored.
	Linguistic demands	This refers to the language of the input and the expected language of the output and can also include reference to the audience or interlocutor where appropriate.
Test administration	Security	Refers to systems that are put in place to ensure the security of the entire administrative process.
	Physical organisation	Refers to room setup etc.
	Uniformity	Systems to ensure that all administrations of the test are the same.
The scoring system		
	Theoretical fit	The way in which test performance is assessed must fit with the conceptualisation of the ability being tested. This goes beyond a key or rating scale to inform the philosophical underpinning of rater and examiner selection, training and monitoring.
	Accuracy of decisions	Encapsulates the old idea of reliability, though broadening this to include all aspects of the psychometric functioning of a test.
	Value of decisions	This relates to things like criterion-related evidence such as comparison with measures such as teacher estimates, other test scores, performance standards such as the Common European Framework of Reference (CEFR).

Table 2: Model usage (summarised from O'Sullivan and Weir, 2011)

Context	Purpose	Source
Cambridge ESOL	Test review	Writing - Shaw and Weir (2007) Reading - Khalifa and Weir (2009) Speaking - Taylor (2011)
	Test revision	Cambridge English: First for Schools (FCE) and Cambridge English: Advanced (CAE) - ffrench and Gutch (2006) - Hawkey (2009)
	Test history	Construct representation - Weir (forthcoming)
	Test validation and development	International Legal English Certificate - Corkhill and Robinson (2006) - Thighe (2006) - Green (2009) - Galaczi and Vidakovic (2010) The Teacher's Knowledge Test - Ashton and Khalifa (2005) - Harrison (2007) - Novakovi´c (2006) Asset languages - Ashton (2006) - Jones (2005) Business English Certificates (BEC) and BULATS - O'Sullivan (2006)
Other contexts of use	Basis of development	Generic English for Specific Purposes (ESP) test specifications - QALSPELL (2004) Zayed University preparatory English programme assessment system - O'Sullivan (2005) EXAVER (Mexico) - Florescano Abad et al. (2011)
	Linking to standards	City and Guilds English for Speakers of other Languages (ESOL) suite - O'Sullivan (2009a, 2009b, 2009c, 2011) COPE (Bilkent University, Ankara) - Kantarcıoğlu (forthcoming) - Kantarcıoğlu et al. (2010)

The real strength of this model of validation is that it comprehensively defines each of its elements with sufficient detail as to make the model operational and while much work is still needed to finalise the model (and in truth it may never be fully finalised), it has already been shown to offer a useful and practical approach to test development and validation. O'Sullivan and Weir (2011) describe in some detail the areas in which the model has already been used; these are summarised in Table 2. The developers of the Aptis system took these elements into account in a number of ways during the development process. This is discussed in the following section.

2.2. Application of the model

The way in which the development team approached the whole process is presented here in terms of the three main elements of the validation model: the population, the test system and the scoring system.

2.2.1. The population

The model of validation was considered from the beginning of the project in the way the expected population for the Aptis system was considered. Since the system is designed for a wide, mainly young adult and adult, population, coming to a clear definition of the population was not feasible. This is, in fact, the case with all international examinations, and even many national examinations which are aimed at a broad audience. For example, an examination in London would need to deal with a potential population from a total of over 230 language backgrounds (UCL, 2012 – website www.ucl.ac.uk/psychlangsci/research/linguistics/dll).

The decision to limit the use of the Aptis system with learners over 15 was based on our experience with the International Language Assessment (ILA). Here, it was found that younger learners did not appear to be compromised by the test, though at lower levels there did appear to be an issue with some learners. The contents of the test are designed to be used with learners irrespective of gender, language or cultural background and are monitored at the production stage for these variables.

2.2.2. The test system

The underlying validation model suggests a different set of parameters related to the tasks used in a test. These parameters include a series related to performance, linguistic demands (input and output) and administration. In this section, we present the approach taken for each of the papers developed to date.

It should be stressed at this point that Aptis is a dynamic system, which has been designed to change over time to reflect current research in second language acquisition, applied linguistics, and assessment. This philosophy has been reflected in the whole approach taken, from the planning of each paper (with a variety of tasks for each skill area from which the user can potentially choose to create an assessment appropriate to his/her context), to the design and specification stage (where Aptis moved away from the traditional paper-based specification document, which tends to fossilise probably due to the 'permanent' nature of the document, and devised a revolutionary new interactive system based on a wiki), to the advanced and flexible delivery system selected by Aptis (a unique and fully integrated test development, management, delivery and monitoring system).

In the following tables (Table 3 and Table 4) we outline how Aptis deals with the various task parameters as outlined in the validation model. The emphasis is on providing the candidate with a broad range of linguistic experiences during the test in order to ensure that the range of evidence built up about the individual is broad enough (and consistent enough) to ensure an accurate and reliable estimate of their ability in each of the Aptis system papers is achieved.

Table 3: Task setting parameters

The test system	
Purpose	In each of the papers in the Aptis system, candidates are offered a wide variety of tasks, each with specifically defined purposes. The rationale behind this approach is to ensure as broad a coverage of the underlying construct as possible and to ensure that candidates are encouraged to set goals from the beginning of each task that reflect those expected by the development team. The flexibility of the Aptis approach means the British Council will be in a position to work with clients to localise (i.e. make appropriate to the particular context and domain of language uses) the test, thus ensuring it will meet the expectations and requirements of the client while maintaining its internal integrity (from a content and a measurement perspective).
Response format	In the same way the different items and tasks have a variety of purposes, they also contain a range of response formats, from multiple choice to matching in the knowledge and receptive skills papers, to structured and open responses in the productive skills papers. This commitment to offering a wide variety of task and item formats reduces the potential for any format-related bias (either positive or negative).
Known criteria	In order to ensure that all candidates set similar goals with regard to their expected responses, the assessment criteria for all tasks and items are made clear both within the test papers and on the Aptis website. It is also the case that the assessment criteria were very carefully considered by the development team in the early stages of the process to ensure that they reflect the underlying knowledge and ability being assessed in each paper. This link is recognised by Weir (2005) and O'Sullivan and Weir (2011) as being critical to the validity of the test.
Weighting	All items are equally weighted in each paper and this information is made clear to the candidates both within the paper and on the Aptis website. This is done to ensure that candidates are all equally informed as the expectations of the developers (and therefore do not spend more time than intended on particular aspects of the test).
Order of items	While the papers are set out in a particular order, the candidate is free to respond in any order, with the exception of the speaking and the listening papers.
Time constraints	Candidates are allowed a limited amount of pre-performance preparation time for both writing and speaking (the time is built into the response times). In addition to this, the time allowed for responding to items and tasks is carefully controlled to ensure a similar test experience for all candidates. In fact, all timings are automatically gathered and will be used by the Aptis research team to study specific aspects of the test papers.

Table 4: Task demands parameters

The test system	
This relates to the language of the INPUT and of the EXPECTED OUTPUT	
Channel	In terms of input, this can be written, visual (photo, artwork, etc.), graphical (charts, tables, etc.) or aural (input from examiner, recorded medium, etc.). Output depends on the ability being tested, though candidates will use different channels depending on the response format (see above).
Discourse mode	Includes the categories of genre, rhetorical task and patterns of exposition and is reflected in the input (in the receptive skills papers) and in the output (in the productive skills papers). A good example of how we approach this is in Task 4 of the writing paper. Here, the candidate is asked to respond to a dilemma in a number of ways, for example in an email to a friend and in a letter of complaint to a person in authority (thus identifying a candidate's ability to recognise register in their written response).
Text length	The amount of input/output depends on the paper, with texts of up to 750 words in the reading and expected outputs ranging from 30 seconds to two minutes in the speaking paper and from 20 to 150 words and writing paper.
Writer/speaker relationship	Setting up different relationships can impact on performance (see Porter and O'Sullivan, 1999). Therefore, where appropriate throughout the Aptis system, efforts have been made to specify (usually within the purpose of the task) the intended interlocutor or audience. An example of specifying different audiences is outlined in the discourse mode section above.
Nature of information	Since more concrete topics/inputs are less difficult to respond to than more abstract ones and the intended candidature is generally at the intermediate or lower levels, the decision was made early in the development process to use texts that were more concrete in nature. However, in the productive tasks, we have deliberately ensured that the expected output will vary from the concrete to the more abstract. An example of this is where three questions are asked (e.g. in relation to a photograph), the first will typically ask for a basic description, the second will ask the candidate to relate the input to their own experience, while the third will ask the candidate to speculate or offer an opinion. Thus the cognitive load of the task gradually increases and with it the difficulty of the task.
Topic familiarity	Greater topic familiarity tends to result in superior performance. Therefore (for a similar reason to that outlined above) it was decided that topics would be used which were likely to be known to the candidates. Difficulty would be manipulated by increasing the cognitive load (through increasing or reducing parameters such as planning time, degree of concreteness, stipulation of audience and expected purpose and amount of output.
Linguistic	
Lexical range	These relate to the language of the input (usually expected to be set at a level below that of the expected output) and to the language of the expected output. For the Aptis papers, the developers have described this language in terms of the British Council/EQUALS Inventory (2011) and the CEFR. The input for all papers is strictly controlled by the development team. All written input is expected to reflect a specified cohesive, lexical and syntactic profile established using online resources such as VocabProfile and Coh-Metrix.
Structural range	
Functional range	

Table 5: Scoring validity

Scoring validity	
Criteria/Rating Scale	The criteria for both the speaking and writing papers are based on the same socio-cognitive theory of language that drives our understanding of language and the tasks we use to elicit that language. Two views of language assessment are reflected in the criteria for the two productive skills. In the writing paper, a holistic approach is taken, in which the reader is expected to apply a single descriptive scale when forming a judgement on the written work. This was done both to reflect the nature of the writing tasks (candidates are asked to perform a series of relatively short tasks, often reacting to brief written prompts) and to facilitate ease of marking. In the speaking paper, where the examiner has some time to reflect on the language of the response, the tasks are assessed using a number of criteria in a series of task-specific scales. The descriptors of ability contained vary with the different tasks to reflect the expected language of the responses.
Rating procedures	
Training	All raters are trained using a standardised system devised for the Aptis project. Raters are also expected to pass an accreditation test at the end of the training event. In-service training is also provided for raters who do not meet the expected level of accuracy or consistency. This training consists of online 'refresher' packs towards which raters are directed if they fail to mark a number of pre-set scripts as expected.
Standardisation	Raters are standardised during the training event, to recognise the type of language expected at the different CEFR levels. In addition, we systematically include control items (CIs), i.e. spoken or written performances which have been scored by a number of senior examiners, in the work to be rated by the examiners.
Conditions	As Aptis uses a secure online rating system, all work is undertaken at the convenience of the raters. This means that the work can be undertaken at a time and place that best suits the rater and results in raters defining their own conditions. Together with the Aptis guide to best practice, this rater-centred format contributes to rater consistency.
Moderation	Moderation is managed automatically within the rating system by placing pre-scored performances in the batches to be rated by each examiner. Failure to rate a set number of these within a given tolerance means that the examiner is automatically offered an in-service training package. Upon successful completion of this package, rating can resume. In addition, all rater data will be explored using many-facet Rasch analysis to highlight inconsistency, harshness and other potentially problematic rater behaviours.
Analysis	Statistical analysis of all rater performances will ensure that individual candidates will not lose out in situations where examiners are either too harsh/lenient or are not behaving in a consistent manner. This is the aspect of Validity that is traditionally seen as reliability (i.e. the reliability of the scoring, or rating, system).
Grading and awarding	Grading and awarding is automated within the Aptis system. This is facilitated by using performance on the core (language knowledge) paper to inform borderline decisions. However, routine analysis of test data will be undertaken to ensure accuracy of all decisions made.

2.2.3. The scoring system

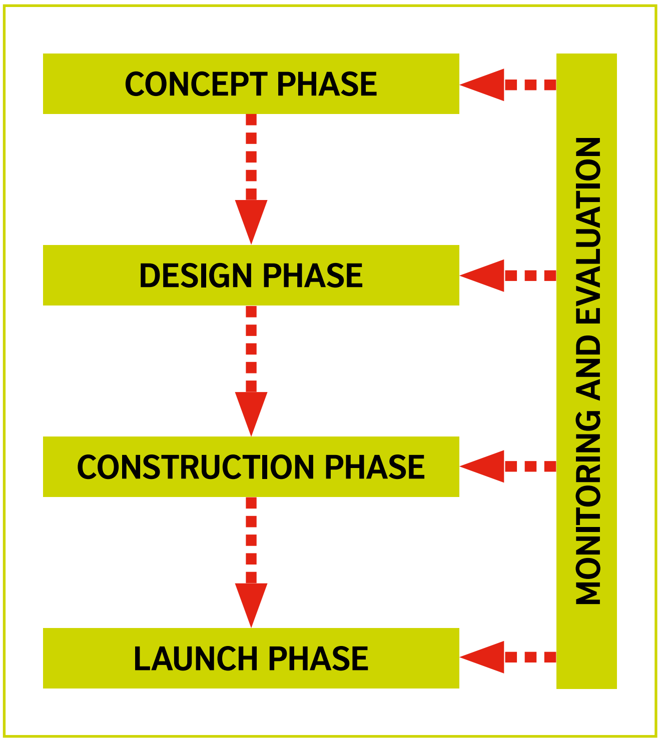
It is important that the scoring system 'fits' with the other elements of the test, otherwise inconsistencies will emerge.

One example of this is where the rating criteria are not suitable for use with a particular task because the expected language (as reflected in the criteria) has not been elicited so the criteria are redundant. Great care has therefore been taken in the Aptis papers for written and spoken production to create rating scales that are designed specifically for use with the tasks created for the test. Table 5 presents an overview of the scoring system.

2.3. Test development model

The Aptis development model is presented in Figure 2. It is important to note that aspects of monitoring and evaluation were applied across all stages of the process. This was done to ensure that the quality of the finished product was as would be expected from a high stakes test (even though the Aptis is not seen as such a test). This commitment to quality marks all aspects of the Aptis system.

Figure 2: The Aptis Development Model



The model is further explored in the following sections.

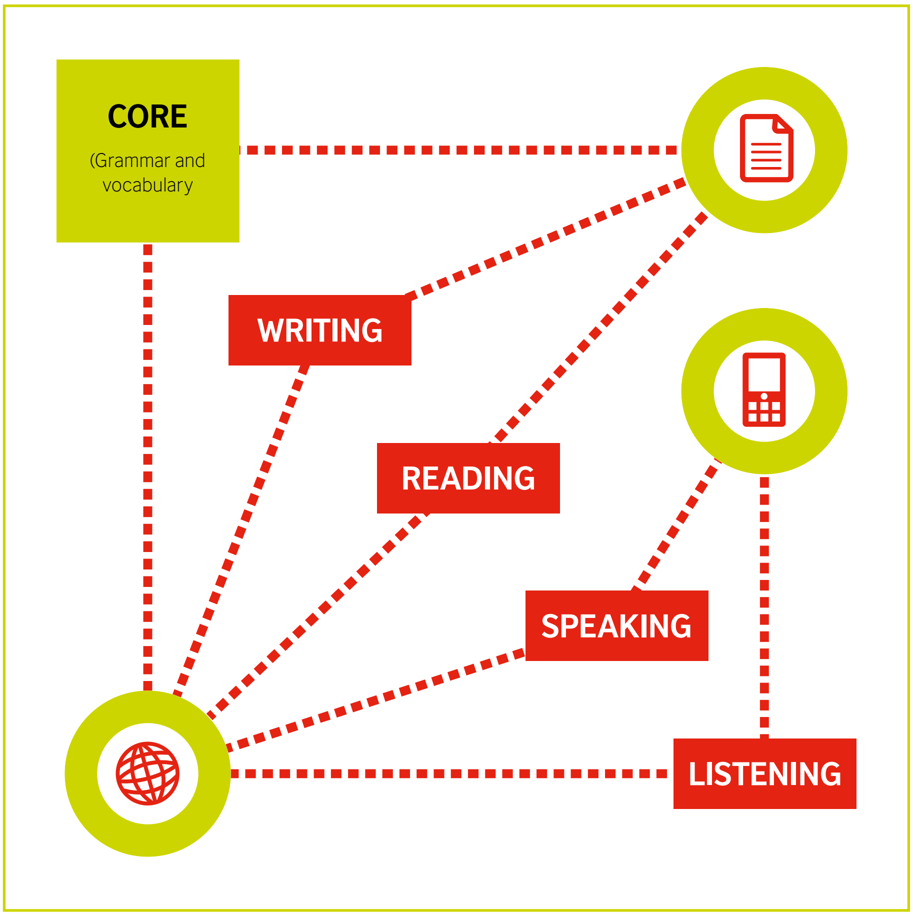
2.4. The concept phase

From the outset, the Aptis system was based on two main ideas, those of flexibility and accessibility. Within these the development team considered areas such as content, structure, delivery mode, reporting options and cost.

In terms of test content, the system was expected to consider flexibility by creating a range of task and item types for each paper. These would then be available to the test producer to create individualised papers for specific clients. The launch version of Aptis will be based around set papers, but the development of additional tasks post-launch will allow for this level of flexibility to be gradually introduced to the system. This will also allow for rapid response to requests from clients for customised (i.e. domain specific) tasks and tests. Accessibility was another consideration in our decision to work towards a variety of task types, as it was considered important that the finished tests reflected the needs of as broad a range of learners as possible.

In order to offer a note of consistency of assessment approach, the developers turned to the research and development work that underpinned the British Council placement test (the International Language Assessment – ILA) project as the basis for the grammar, vocabulary and reading papers. The other papers were developed from scratch.

Figure 3: Aptis delivery modes



To further reflect our two main driving goals (flexibility and accessibility) we conceptualised the test as being flexible in terms of how the different elements (i.e. the skills papers) could be combined in ways that meet the needs of the client. The combination of papers, which are always taken with a core language knowledge element (i.e. grammar and vocabulary) offers significant flexibility and also ensures that tests are more readily accessible, as the client only pays for what is actually needed. The decision to deliver the tests using a variety of delivery modes (pen and paper, computer and phone – see Figure 3) also contribute to flexibility and accessibility.

At the time of launch, all computer delivery options were available, as were the pen and paper core and reading papers.

2.4.1. The design phase

The Aptis papers were designed by a team of British Council teachers and assessment specialists who were first offered an intensive continuing professional development (CPD) programme on the theory and practice of test development. This programme was based around the same model of validation that drives the test to ensure that all project participants were fully cognisant of the underlying theories. Following on from the CPD, a series of workshop/seminars were delivered in the UK to all participants to develop the ideas from the CPD into test tasks for inclusion in the Aptis system. The total number of hours training offered to team members was over 250, far in excess of the industry average. The purpose of this extensive training was to identify individuals who could, in the medium to long term, bring the sort of depth to the British Council's assessment expertise that typified the organisation up to the late 1980s.

The work of this team also took into consideration the earlier ILA project work as well as the British Council EAQUALS Core Inventory (which had also been influenced by the ILA research into grammatical progression) and the Common European Framework of Reference for Languages (CEFR).

The team worked to devise appropriate tasks and to create working specifications (blueprints) for all tasks and items developed. Initially, these specifications were more traditional in form. However, in response to the fear that by simply writing the specification down they might be seen as 'final' we conceived of the idea of creating a wiki. This allowed us to consider the specifications as more of a living text, which could be contributed to by the team. This approach, which is the first time a wiki has been used for test specification, has been extremely successful in ensuring that the approach to writing and developing test tasks has not fossilised, but continues to be fresh and innovative. It also secures the link between item writer behaviour and the specifications, as all item writers are encouraged to work with the wiki, offering advice to others on specific task development and generally sharing their experiences.

Items developed during this phase of the project were trialled with small populations to establish that they were likely to result in the level and type of performances expected.

2.4.2. The construction stage

The construction phase has been based around the development of an item bank of test tasks and items. The training of item writers (up to now these have been the same people who developed the items and wrote the specifications) was described briefly in the previous section, and was primarily focused on aspects of quality. When items are written and have successfully come through a systematic review, they are trialled with a representative population of learners. This trial has two purposes; firstly to ensure that the item is working as predicted, and secondly to estimate its true difficulty. The former is done through analysis of the response data, and the latter is done using item response theory (IRT) based statistics. This type of test data analysis is based on probability theory and offers the test developer an estimate of the probable true difficulty of the item – which, unlike classical test analysis, is not linked to a particular population. This estimate, which is in the form of a number called a logit (pronounced: loh-jit) value, is useful in that it allows for tests to be created to a pre-determined level of difficulty. As such, it is vital for the creation of an item bank.

The populating of the bank of test tasks and items is now well under way, with team members now working with our platform providers (BTL) to input the almost 2,000 items that have already been written into the system.

In addition to the processes described above, we have been working work on the other elements of the system, from examiner training packages to the writing of manuals.

A series of trials was undertaken in spring and early summer of 2012. In the first of these, approximately 250 participants sat for a whole series of Aptis papers. These trials allow us to understand the relative connections between the various papers, particularly important for the language knowledge paper which is seen as the underlying link between all Aptis papers. They also allow us to begin the process of more fully understanding the underlying measurement model of the test, ensuring that significant assumptions (such as item independence) can be supported.

Finally, a major formal trial was undertaken in which almost 900 candidates across the world sat for a variety of Aptis test packages (see TechRep-02 for a detailed report on feedback from stakeholders of this trial). The purpose of this trial was to add further to our understanding of how the various papers were working and to gather feedback from a range of stakeholders on how all aspects of the test, from the visual presentation, to the perception of ease/difficulty, to the administrator and invigilator guidelines. Feedback from the trials allowed us to finalise the working guidelines for the test, though we recognise that experience across the world in delivering Aptis will contribute to the guidelines in the initial year.

2.4.3. The launch phase

Aptis was successfully launched in August 2012. At this point in time, the work of the development team changed to some extent. While we continue to work on the building of the item bank and on the development of new item types, we have instigated a systematic monitoring and evaluation programme. This programme offers regular systematic reviews of the Aptis system to ensure that the different papers are working as intended and that the various guidelines and handbooks are kept up to date. The team will create an annual report on the quality and use of the Aptis system.

The avenues of research and development that form part of our continuing work include:

- developing new task/item types for general use
- developing tasks/items for use in specific domains (e.g. business, education)
- developing tasks/items for use with specific populations (e.g. country or region specific)
- researching aspects of the Aptis system (e.g. rating processes)
- researching aspects of Aptis use (e.g. benchmarking to specific jobs).

To facilitate this research and development work, the British Council has signed a formal memorandum of understanding with the Centre for Research in English Language and Language Assessment (CRELLA), University of Bedfordshire, currently Europe's foremost English language assessment research centre. In addition, the British Council has created a new research initiative, called the Aptis Research Awards, which will be launched in autumn 2012.

References

Abad Florescano, A, O'Sullivan, B, Sanchez Chavez, C, Ryan, D E, Zamora Lara, E, Santana Martinez, LA, Gonzalez Macias, MI, Maxwell Hart, M, Grounds, P E, Reidy Ryan, P, Dunne, RA and Romero Barradas, T de E (2011) Developing affordable 'local' tests: the EXAVER project, in Barry O'Sullivan (ed) *Language Testing: Theory and Practice* (pp. 228–243). Oxford: Palgrave.

Ashton, K (2006) Can Do self-assessment: investigating cross-language comparability in reading. *Research Notes* 24, 10–14.

Ashton, M and Khalifa, H (2005) Opening a new door for teachers of English: Cambridge ESOL Teaching Knowledge Test. *Research Notes* 19, 5–7.

Chizek, GJ (2011) *Reconceptualizing Validity and the Place of Consequences*. Paper Presented at the Annual Meeting of the National Council on Measurement in Education New Orleans, LA, April 2011.

Corkhill, D and Robinson, M (2006) Using the global legal community in the development of ILEC. *Research Notes* 25, 10–11.

ffrench, A and Gutch, A (2006) FCE/CAE Modifications: Building the Validity Argument: Application of Weir's Socio-Cognitive framework to FCE and CAE. *Cambridge ESOL internal report*.

Galaczi, ED and Vidakovic, I (2010) Testing Legal English: Insights from the International Legal English Test. *Professional and Academic English*, 35 (March).

Green, A (2009) *A Review of the International Legal English Certificate* (ILEC). Unpublished report submitted to Cambridge ESOL Examinations, April 2009.

Hawkey, RA (2009) *Examining FCE and CAE: Key issues and recurring themes in developing the First Certificate in English and Certificate in Advanced English exams. Studies in Language Testing* 28 Cambridge: Cambridge University Press.

Jones, N (2005) Raising the Languages Ladder: constructing a new framework for accrediting foreign language skills. *Research Notes* 19, 15–19.

Kantarcıoğlu, E (forthcoming). *A Case-Study of the Process of Linking an Institutional English Language Proficiency Test (COPE) for Access to University Study in the Medium Of English to the Common European Framework for Languages: Learning, Teaching and Assessment*. Unpublished PhD thesis, University of Roehampton, London.

Kantarcıoğlu, E, Thomas, C, O'Dwyer, J and O'Sullivan, B (2010). 'The cope linking project: a case study' in Waldemar Martyniuk (ed) *Aligning Tests with the CEFR: Case studies and reflections on the use of the Council of Europe's Draft Manual. Cambridge: Cambridge University Press*

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

Novakovi´c, N (2006) TKT – a year on. *Research Notes* 24, 22–24.

O'Sullivan, B (2006) *Issues in Testing Business English: The BEC Revision Project, Studies in Language Testing* 17. Cambridge: Cambridge University Press.

O'Sullivan, B (2005) *Levels Specification Project Report*. Internal report, Zayed University, United Arab Emirates.

O'Sullivan, B (2009a) *City and Guilds Communicator Level IESOL Examination (B2) CEFR Linking Project Case Study Report*, City and Guilds Research Report. Available online at www.cityandguilds.com/documents/ind_general_learning_esol/CG_Communicator_Report_BOS.pdf

O'Sullivan, B (2009b) *City and Guilds Achiever Level IESOL Examination (B1) CEFR Linking Project Case Study Report*. City and Guilds Research Report.

O'Sullivan, B (2009c) *City and Guilds Expert Level IESOL Examination (C1) CEFR Linking Project Case Study Report*. City & Guilds Research Report.

O'Sullivan, B (2011) Language Testing. In James Simpson (ed) *Routledge Handbook of Applied Linguistics*. Oxford: Routledge.

O'Sullivan, B and Weir, C (2011) Language Testing and Validation, in Barry O'Sullivan (ed) *Language Testing: Theory and Practice* (pp. 13–32). Oxford: Palgrave.

QALSPELL (2004) Quality Assurance in Language for Specific Purposes, Estonia Latvia, and Lithuania. Leodardo da Vinci funded project. Website: www.qalspell.ttu.ee/index.html

Shaw, S and Weir, CJ (2007) *Examining writing: Research and practice in assessing second language writing, Studies in Language Testing 26*. Cambridge: Cambridge University Press and Cambridge ESOL.

Taylor, L (ed) (forthcoming) *Examining Speaking: Research and practice in assessing second language speaking, Studies in Language Testing 30*. Cambridge: UCLES/Cambridge University Press.

Thighe, D (2006) Placing The International Legal English Certificate on The CEFR. *Research Notes* 24, 5–7.

Weir, CJ (2005) *Language Testing and Validation: an evidence-based approach*. Oxford: Palgrave.