BRITISH COUNCIL

ENGLISH LANGUAGE
ASSESSMENT RESEARCH GROUP

APTIS–VSTEP COMPARABILITY STUDY: INVESTIGATING THE USAGE OF TWO EFL TESTS IN THE CONTEXT OF HIGHER EDUCATION IN VIETNAM

VS/2018/001

**Jamie Dunlea, Richard Spiby,
Thi Ngoc Quynh Nguyen, Thi Quynh Yen Nguyen,
Thi Mai Huu Nguyen, Thi Phuong Thao Nguyen,
Ha Lam Thuy Thai and Bui Thien Sao**

BRITISH COUNCIL VALIDATION SERIES
SERIES EDITOR: VIVIEN BERRY
ISSN 2398-7979

ABSTRACT

This study is the second in the British Council Validation Series to focus on investigating test comparability. These studies aim to contribute to the theoretical framework for carrying out such comparability studies by using the socio-cognitive model for language test validation to design a multi-method data collection and analysis approach. In particular, an important part of both studies has been the use of detailed content analysis approaches to build a construct definition to inform interpretation of other sources of data, particularly quantitative data analysis evidence. This research focus is also informed by, and contributes to, the importance of localisation within the theoretical framework of test development and validation which has been at the centre of the Aptis test development approach from the beginning (see O'Sullivan, 2015a, O'Sullivan and Dunlea, 2015).

This particular study, as with Wu et al. (2016), reports on a comparability study of two EFL proficiency tests which use an international proficiency framework, the Common European Framework of Reference for Languages (CEFR), as an important source of feedback for test takers. The two tests in this study are VSTEP, a pen-and-paper test in Vietnam targeting CEFR levels B1 to C1, and APTIS, an international computer-based test targeting CEFR levels A1 to B2. The VSTEP is recognised by universities in Vietnam as certification of English proficiency for the purpose of meeting graduation requirements stipulated by the Ministry of Education. Aptis is used for a range of purposes in international settings, including by higher education institutions in EFL contexts.

This study reports on both a smaller scale pilot phase at one university to trial the methodology before describing the main phase in which over 400 test takers at three universities in different regions took both tests. The socio-cognitive model for language test development and validation was used to design a multi-method approach. Statistical analysis of test scores includes factor analysis, as well as a concurrent Rasch analysis to place test items from both tests on a common scale. In addition, a comprehensive pro forma, utilising categories drawn from the socio-cognitive model and the growing body of CEFR alignment studies, was developed to evaluate the constructs targeted by both tests. Questionnaire data from test takers also offers interesting insights into the attitudes of the university students regarding such aspects as differences in delivery mode for the productive skills.

Authors

Jamie Dunlea is a Senior Researcher for the Language Assessment Research Group at the British Council, based in London. He has worked on a range of language test development and validation projects with the British Council, as well as collaborating on projects with researchers, organisations, and ministries internationally. Jamie joined the British Council in 2013, and was previously Chief Researcher at the Eiken Foundation of Japan, a not-for-profit organisation which develops and administers EFL examinations in Japan.

Richard Spiby has been the Test Development Researcher for the receptive skills (reading and listening, together with grammar and vocabulary) at the British Council since June 2016. His main responsibilities involve analysing operational data, and revising and developing the receptive skills components of the Aptis test. Richard has previously worked in the UK and Turkey, mainly in the university sector, in test production, management and research. His particular areas of interest are cognitive processing in language assessment, strategy use in reading and listening, and methods of testing vocabulary.

Nguyen Thi Ngoc Quynh (Quynh Nguyen) holds a PhD in Applied Linguistics from the University of Melbourne, Australia. She is currently the Director of the Center for Language Testing and Assessment at the University of Languages and International Studies, Vietnam National University Hanoi. She plays a leading role in multiple institutional and national projects on language assessment and teacher development. Her research interests are second language education and assessment, teacher development, and bilingual education.

Nguyen Thi Mai Huu (Huu Nguyen) is the Director of the National Foreign Languages Project, Ministry of Education and Training of Vietnam. She is currently a PhD candidate in Applied Linguistics at the University of Languages and International Studies, Vietnam National University Hanoi. Her research interests include language test development and validation, test specifications, classroom-based assessment, and assessment for learning. She was involved in the development of the Vietnam Standardised Test of English Proficiency (VSTEP) in 2015, which is Vietnam's national test.

Nguyen Thi Quynh Yen (Yen Nguyen) has taught English for 15 years in the University of Languages and International Studies, Vietnam National University Hanoi (ULIS-VNU). She holds an MA in Teaching English as a Second Language and is now a PhD candidate in English Language Assessment. She is the Deputy Director of the Center for Language Testing and Assessment, ULIS-VNU. She has been actively involved in English teacher training programs in Vietnam. Her research interests include English linguistics, teaching methodology and language assessment.

Nguyen Thi Phuong Thao (Thao Nguyen) is a test developer and researcher at the Center for Languages Testing and Assessment, University of Languages and International Studies, Vietnam National University Hanoi. She earned a Master's degree in English Language Teaching from the University of Southampton, UK in 2012. She has participated in a number of research projects on VSTEP test design and development, the ULIS university entrance exam, and examiner training. Her research interests include teacher education, professional development, testing and assessment.

Thai Ha Lam Thuy (Thuy Thai) has been a Test Development researcher for reading skills of the VSTEP test since 2016. She has been in charge of analysing operational test data and developing and revising VSTEP reading items. She has also been involved in developing and delivering rater training programs for the VSTEP speaking component. She is currently pursuing her PhD degree in Language Testing and Assessment at the University of Huddersfield in the UK. Her areas of interest are rater cognition, and the development process for raters' expertise.

Bui Thien Sao (Sao Bui) works as a test developer at the Center for Language Testing and Assessment, University of Languages and International Studies, Vietnam National University. She received her Master's degree in English Education from Korea University, Seoul, Korea. Her research interests include language testing and assessment, reading comprehension, and learner differences.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 Overview

This paper reports on a comprehensive test comparability study carried out over two years to investigate the similarities and differences between two English as a Foreign Language (EFL) tests in the context of Vietnam. One of these tests, VSTEP, was developed locally as a national standardised proficiency test for use in Vietnam. The other, Aptis, was developed by the British Council as a test of general English proficiency offered internationally across a range of contexts. The project was carried out as a collaborative research project between the Assessment Research Group at the British Council, the British Council Vietnam exams team, and the University of Languages and International Studies, Vietnam National University, Hanoi (ULIS). The study was supported with funding from the National Foreign Languages 2020 Project. The project grew out of ongoing collaboration and technical and information exchanges between the British Council in Vietnam and Project 2020 to support the reform of English language education and assessment in Vietnam.

The project took place over two years, from 2016 to 2017, against the backdrop of dynamic changes to the English education and assessment landscape in Vietnam. University graduates in Vietnam are required to demonstrate pre-set levels of English proficiency in order to graduate. For non-English majors, this is set at B1 on the Common European Framework of Reference for Languages (CEFR), or the equivalent on Vietnam's six-level framework of foreign language proficiency, which was adapted from the CEFR (referred to as the CEFR-VN for short). Students can take suitable proficiency tests, or tests produced by their institutions, to demonstrate this requirement. This has led to wide variability in the quality of assessments, and the VSTEP was developed as an important step in helping to provide standardised, national measures of proficiency which could be used to raise standards. The study was carried out by a project team, with members from the Assessment Research Group and ULIS, with the British Council Vietnam team supporting the project logistically for administration of Aptis tests, etc.

This report will explain the rationale for carrying out the study, provide an overview of the methodology employed, including the approach to content analysis and construct definition, and describe the results from the study, including scoring comparisons and questionnaire data obtained from students in the pilot and main studies.

1.2 Aptis

The Aptis test system is not just a single test. The system was developed at the British Council to provide a coherent, theory-based approach to language test development which could be applied a range of uses and test user groups. There are a number of variants live within the system. The variant used for this study is Aptis General, which was the first variant developed within the system. Aptis General was first administered in 2012. It is a computer-based test of general English proficiency for adult test takers (16 years and older). It comprises five components: a 'core' grammar and vocabulary component, as well as reading, listening, speaking and writing components. The test was designed to prioritise flexibility, and the concept of localisation has been an important part of the theoretical model. A model of localisation has been developed to provide a framework for adaptations. At its most basic, test users can choose a standard variant but can choose components within that variant that are most useful for their needs and resources, for example, selecting reading and listening, but forgoing the productive skills components (the core component is a compulsory element of the test).

In terms of feedback for test users, Aptis provides scale scores for all components, on a scale of 0–50. For the four skills components of reading, listening, writing, and speaking, a CEFR level is also provided. When all four skills have been included by a test user, an overall CEFR level averaging the CEFR levels across the skills profile is also provided. The two types of feedback are in line with the concept of flexibility and the wide range of test users with different needs. The scale scores provide the opportunity to look at more fine-grained, incremental changes in proficiency. The CEFR levels provide a more criterion-referenced focus for interpreting proficiency levels. A detailed overview of the test system and Aptis general in particular is provided in the *Aptis General Technical Manual* (O'Sullivan and Dunlea, 2015). Information on the test, including examples of item types and a range of research publications, is also available online (see http://www.britishcouncil.org/exam/aptis/research).

1.3 Vietnamese Standardised Test of English Proficiency (VSTEP)

VSTEP stands for Vietnamese Standardised Test of English Proficiency. This is the first standardised English proficiency test in Vietnam. The test specifications and format were developed by language testing experts from the University of Languages and International Studies, Vietnam National University, Hanoi (ULIS). It was released nationally under the auspices of the Ministry of Education and Training (MOET) on 11 March 2015. VSTEP targets adult test takers for a range of general English proficiency purposes in Vietnam, 18 years or older. However, one of its primary uses is in the proof of English ability for university graduates, with non-English majors required to demonstrate a B1 level of proficiency, as noted above. English majors are required to demonstrate higher levels of proficiency, and so the test targets levels B1 to C1 on the Common European Framework of Reference for Languages (CEFR).

The test was developed with three main purposes in mind.

1. To build and conduct an English language proficiency test for Vietnamese learners.
2. To assess the English language proficiency of Vietnamese learners under CEFR-VN standards (the version of the CEFR adapted for use within the context of Vietnam).
3. To establish and prove the national testing capability of Vietnam.

VSTEP is a pen-and-paper four-skills test testing reading, listening, speaking and writing. The speaking test is a face-to-face interview format test with a trained examiner. VSTEP provides feedback to students in terms of their overall proficiency level. Raw scores are converted to band scores on a 0–10 scale. The test further reports whether test takers have achieved a level relevant to CEFR B1, B2, C1, or are below B1 (Dunlea et al., 2016). (The results are reported on a six-level numerical scale used as a local adaptation of the CEFR, CEFR-VN, with 3 being equivalent to B1, 4 to B2, and 5 to C1.) The focus on proficiency levels B1 and above reflects the usages for which the test was developed, including as noted above, being used as proof of attainment of proficiency levels laid out as requirements for university graduation by MOET.

The VSTEP project has led to a number of collaborative interactions between the ULIS project group and international researchers, in addition to this project with the British Council. From this perspective, apart from the development of the test itself, the project has provided an important source of professionalisation for language testing researchers in Vietnam.

1.4 Rationale

The project presented the participating organisations with the opportunity to contribute to both the body of validation evidence for their own tests, as well as contribute to the wider field by elaborating theoretically sound and practically achievable methods for investigating test comparability.

From the first perspective, the focus and goals were inward looking, and concerned with investigating the comparability of two tests being used for similar purposes in the context of Vietnam. From this perspective, the goals for both organisations were largely instrumental. Aptis is an international test of general English proficiency which has been used increasingly within higher education in EFL contexts, including Vietnam. VSTEP is a national test designed to suit the English language use context in Vietnam. While both VSTEP and Aptis are standardised tests, they have several variants. For Aptis, these variants are aimed at different test user groups. For VSTEP, they are delivered at different levels (for levels 3–5 (or B1–C1), and level 2 (or A2-based) to date). Thus, from their outset, the methodological framework for both these tests has emphasised the concept of localisation and investigating local appropriateness.

This study provided the opportunity to investigate the use of Aptis with university students, not just from a scoring perspective, but also in relation to their attitudes and perceptions, for example, regarding the computer-delivered formats for speaking and writing. For VSTEP, the study provided the opportunity to investigate further how the VSTEP design would play out in terms of international benchmarks of proficiency. For VSTEP, currently delivered in pen-and-paper mode, the study also offered insights into student preferences on delivery mode, useful for informing decisions on the possibility of introducing computer-based formats for each and every skill in the future. As both tests claim alignment with the Common European Framework of Reference (CEFR), and both tests have carried out studies to support their claims, the CEFR provided a focal point for comparing the performance of students and if the two tests would classify students in broadly similar ways according to this proficiency framework.

At the same time, the study has a wider goal to contribute to the methodological framework within which test comparability studies, particularly in relation to the CEFR, can be carried out. In particular, through the use of the socio-cognitive model for language test development and validation, the study explicitly aimed to emphasise the importance of going beyond score comparisons and including a strong element of content analysis to inform construct definitions for the two tests.

2. LITERATURE REVIEW

The following short review provides an outline of studies most relevant to the mixed-methods approach employed by this study. The present study builds on a methodological framework piloted in Wu, Yeh, Dunlea, and Spiby (2016), which is described further below. A comprehensive overview of test comparability approaches is beyond the scope of this report. What we aim to provide is a summary of the studies most relevant to the approach taken, including a description of the socio-cognitive model for language test development and validation, which informed the study design, and how that model, and the approach taken here, are also related to the literature on linking to the CEFR.

A seminal project in language test comparability studies is the systematic approach to investigating similarities between two international language test batteries often referred to as the Cambridge TOEFL Comparability Study (CTCS), and reported in Bachman, Davidson, Ryan and Choi (1995). This study utilised a range of both quantitative and qualitative methods to compare the characteristics of typical test takers, test uses, test content and quantitative measures of test performance, such as means, correlations and Exploratory Factor Analysis (EFA) to explore the factor structure of the test batteries. The authors utilised content analysis templates based on Bachman's 1990 model of

Communicative Language Ability and Test Method Facets to analyse content. Although the authors set out with a hypothesis that the very different measurement traditions from which the tests were developed would result in "clear and striking differences", both in expert judgment of test content and test performance (Davidson and Bachman, 1990:28), in fact, they found a substantial amount of overlap. While the study has faced some criticism over limitations in sample and study design (e.g. Brennan, 1989), it remains an important touchstone in comparability studies. Indeed, Davidson and Bachman (1990:42), suggest that the study is useful as "an example of cross-national language testing research – possibly even as a research model to be replicated". At the same time, they recognised its limitations and made suggestions for "how the CTCS design could be altered for future international language testing research".

Two important features differentiate the CTCS from the present study. Firstly, the authors were explicitly aiming to investigate tests used in different national contexts, with possibly very different test populations, as well as test uses and measurement cultures. Secondly, there was no common external framework of proficiency to which both tests claimed alignment, making the comparison of performance on the tests, and interpretations of proficiency based on results for the same test takers on the two tests difficult.

A major development in the decades since the publication of Bachman et al. (1995) has been the introduction and rapid uptake of the CEFR. Published in 2001 by the Council of Europe following a 10-year development period, it has been rapidly adopted both inside and outside Europe as an external framework of proficiency. Serious limitations have been recognised in the use of the CEFR in test development, as well as problems stemming from the over-simplistic adoption of the six broad proficiency levels without sufficient attention to local contexts (for example, Fulcher, 2004; Alderson et al., 2006; Davidson and Fulcher, 2007; Weir, 2005a). Nonetheless, as Alderson (2005) notes, the rapid adoption of the CEFR underscores its utility in providing researchers and educators with a common set of terms and key proficiency benchmarks which can facilitate the comparison of language examinations across contexts. This is, in fact, in line with the original intentions of the CEFR, as noted by North, Martyniuk and Pantheir (2010:13), who note that it needs to be "applied and interpreted with regard to each specific educational context."

In conjunction with the rapid uptake of the CEFR, a large body of literature has been produced on demonstrating alignment of tests with the CEFR in a theoretically sound, evidence-based way. The Council of Europe has produced a number of documents including the *Manual for Linking Exams to the CEFR* (2009). The Manual, much like the CEFR itself, is not intended to be prescriptive or promote only one approach. It does, however, outline a number of stages that a test developer wishing to demonstrate a link to the CEFR needs to address, including *familiarisation*, *specification*, *standardisation* and *standard setting*. Although the Manual places great importance on standard setting, it does not identify any one standard-setting method as superior. Rather, it is the systematic collection and documentation of the method used which is emphasised. Similarly, the Manual does not specify any particular validity model, and makes clear that linking to the CEFR is not in, and of, itself a form of test validation. Preliminary steps to linking include a thorough validation argument for the test itself, establishing a justification for the appropriateness of the uses and interpretations made of the test. Additionally, the Specification stage described in the Manual requires the clear analysis and comparison of skills and abilities targeted by the test and skills and abilities referred to in the CEFR illustrative scales. In other words, establishing a construct definition for the test and explicitly outlining how this is relevant to the construct of proficiency in the CEFR is seen as an important *a priori* step before standard setting.

To further facilitate the specification stage, and also help address the shortcomings of the CEFR for test development, Alderson et al. (2006) developed a series of content analysis grids for evaluating and describing reading and listening tests, and these have now been included in the Manual. The grids have also been adapted to inform test specifications and test analysis projects undertaken under the methodological umbrella of the socio-cognitive model for language test development and validation, which will be discussed next.

Although the Manual outlines a methodological framework for demonstrating evidence to support the link of *a test* to the CEFR, rather than comparing two tests (as in this study), relevant to this study is the importance it places on establishing construct comparability, not just statistical comparability. O'Sullivan (2017) in a keynote address at the *New Directions in English Language Assessment Conference* has further emphasised the importance of providing a clear construct definition and demonstrating how that definition is relevant to an external proficiency framework before linking to that framework. This suggestion, as noted, is equally relevant to studies investigating the comparability of separate tests, such as this study.

In order to drive the systematic collection of both qualitative and quantitative data for the adequate comparison of the tests in this study, including content analysis for construct definition, a suitable validation model was required. The Aptis test was developed using the socio-cognitive model of language test development and validation (O'Sullivan and Dunlea, 2015; O'Sullivan, 2015a), and this model was selected as the methodological framework for the study. The model was first fully outlined in Weir (2005b) and later elaborated and modified in O'Sullivan and Weir (2011). O'Sullivan further described adaptations of the model and its application to the development of Aptis (O'Sullivan, 2011, 2012, 2015a). The model builds on developments in validity theory over the last three decades following Messick's seminal 1989 description of validity as a unitary concept. While argument-based approaches to validation (Kane, 1992, 2002, 2013; Chapelle et al., 2008, 2010) have received a great deal of attention over the proceeding decades, Dunlea (2016a) suggests that the socio-cognitive model and argument-based approaches are not, in fact, in opposition, and provide potentially complementary approaches.

Dunlea (2016a) further suggests that an important contribution of the socio-cognitive model is its specificity in terms of categories of criterial features. While argument-based approaches have explicitly shied away from providing such concrete taxonomies of evidence (Chapelle et al., 2010), the socio-cognitive model has iteratively built a list of relevant categories of evidence that can drive data collection. Dunlea (2016a) has suggested that these taxonomies can be seen as building on the six aspects of validity described by Messick (1989, 1995, 1996), and which provide a concrete framework within which evidence collected across tests can be compared and evaluated using a common set of categories. It is this concrete application to test development which O'Sullivan and Weir (2011) also saw as a major contribution of the model. The content analysis grids developed by Alderson et al. (2006) have been adapted to help inform these taxonomies of criterial features in many of the applications of the studies described below, and thus provide an important potential link to CEFR linking studies employing the Manual.

Applications of the model include a number of test validation studies to comprehensively describe features of Cambridge Main Suite tests (Shaw and Weir, 2007; Khalifa and Weir, 2009; Taylor, 2012; Geranpayeh and Taylor, 2013; Weir et al., 2013), and the EIKEN tests in Japan (Dunlea, 2016a). The model has also been used to drive studies linking tests to the CEFR (O'Sullivan, 2008, 2010; Wu, 2014), and to provide the design and validation framework for test development of tests designed from the outset to incorporate the CEFR to aid in criterion-referenced proficiency description and the interpretation of test performance. These projects include the TEAP test in Japan (Nakatsuhara, 2014; Taylor, 2014; Weir, 2014) and the Aptis test (O'Sullivan, 2015; O'Sullivan and Dunlea, 2015). In the case of Aptis and TEAP, an important development was the use of the taxonomies of criterial features specified in the model and operationalised in various validation studies to develop detailed test task specification tables. The criterial features for validation were thus adapted and used to provide concrete features of task specification to serve as blueprints for item writing and quality assurance. This approach was also used as the framework for developing a set of test analysis templates for comparing proficiency tests in order to evaluate their suitability for use as language requirements for medical professionals (Chan and Taylor, 2015).

The present study builds on this background and, in particular, a framework for investigating test comparability piloted in Taiwan and reported in Wu et al. (2016). That study looked at the comparability of Aptis and the General English Proficiency Test (GEPT) developed by the Language Training and Testing Centre (LTTC), and also employed a mixed-methods approach to collect both qualitative and quantitative data from a number of perspectives, including a detailed content analysis to inform construct definition for the two tests. These two tests also had established links to the CEFR, which enabled a comparison of how the two tests allocated the same test takers to CEFR levels, facilitating an evaluation whether the two tests indeed had a similar interpretation of the CEFR levels. The study was informed by Wu's (2014) use of the socio-cognitive model to develop content analysis templates for the GEPT, and also made use of the approach to test task specification employed by Aptis.

As discussed above, before considering whether to engage in a comparison study, as with linking a test to the CEFR, it is first necessary to consider the validation evidence supporting the technical quality of the tests themselves, for the uses and interpretations they are intended to serve. As noted above, a full description of the technical properties of the Aptis General, the standard variant of the Aptis system used in this study, is contained within the Technical Manual (O'Sullivan and Dunlea, 2015). As well as presenting technical performance statistics, descriptions of the target language use domain, scoring and feedback, the Technical Manual also details test and task specifications and the rating scales employed by raters. An annual technical update which summarises key statistics, such as reliability and standard error of measurement for global operational test data, is also published online (Aptis Technical Update: http://www.britishcouncil.org/exam/aptis/research/publications/technical-report). A description of how the socio-cognitive model was used in the design and development of Aptis is available in O'Sullivan (2015a), and the process of linking to the CEFR is available in O'Sullivan (2015b). At the time of writing, there were also 22 additional technical reports and validation studies, including 15 produced by external researchers through the Assessment Research Awards and Grants scheme published online (see http://www.britishcouncil.org/exam/aptis/research/publications).

The development of the VSTEP has been a major undertaking, and from the outset, was intended to not only develop the first nationally available standardised proficiency test produced in Vietnam, but also to build language testing expertise and capability (Dunlea et al., 2016; Nguyen and Do, 2015). Less has been published in English on the development and validation of the test, with documentation remaining largely as internal reports. However, the results of ongoing collaborations with international researchers have been presented at a number of international, peer-reviewed conferences, including: at the *2015 Language Testing Research Colloquium* in Toronto (Tran, Nguyen, Dang, Nguyen, Nguyen, Huynh, Do, Nguyen, Davidson, 2015) — which included information on the alignment of the test to the CEFR; at the *New Directions in English Language Assessment Conference* in Hanoi (Carr, Nguyen, Nguyen, Nguyen, Nguyen, 2016); and at the *Fourth Asian Association for Language Assessment Conference* in Taipei (Nguyen, Nguyen, Nguyen, Thai, Bui and Carr, 2017). These presentations offer insights into the available validation evidence for the test, and provided some confidence for the project team in initial discussions about the viability of a comparability study for VSTEP and Aptis. Nonetheless, the need for more published documentation, in both Vietnamese and English, is taken up in the discussion in Section 5.

3. METHODOLOGY

3.1 Overview

The methodology and study design have been driven by the socio-cognitive model introduced above in the literature review. Below, we describe the categories of evidence included in the model, and how the data collection in the study design relates to these categories.

3.2 The socio-cognitive model

In its initial formulation in Weir (2005b), the model contained five aspects of validity essential for collecting evidence which would support a balanced, coherent, and comprehensive validity argument in support of the uses and interpretations of a test: content validity evidence; cognitive validity evidence; scoring validity evidence; criterion related validity evidence; and consequences and impact validity evidence. Dunlea (2016a) has shown how these five categories of evidence overlap with the six aspects of validity evidence which Messick suggested were necessary, and sufficient, to ensure that "the theoretical rationale or persuasive argument linking the evidence to the inferences drawn touches the important bases" in a comprehensive validity argument to justify the uses and interpretations of a test (Messick, 1995:747).

The explicit incorporation of the cognitive demands posed by test tasks into the validity evidence framework has been a major contribution of the model. In validating a test, it is essential to establish the cognitive profile of the test tasks, elaborating the cognitive processes that will be elicited when test takers engage in a task. This then enables validation through a comparison of whether these processes resemble similar processes elicited when language users engage in real-life language use tasks in the TLU. This has led to the inclusion of intended task cognitive profiles in task specifications where the model has been used to drive test design. For example, the Khalifa and Weir model of the cognitive processes involved in reading has been incorporated into reading task specifications for Aptis, and a cognitive processing model for listening based on Field (2013) in the specifications for listening tasks (see task specifications in O'Sullivan and Dunlea, 2015). This has enhanced the comparability of tasks intended to target a similar level of difficulty by ensuring that the cognitive demands are comparable. At the same time, it has facilitated external validation studies to validate whether the intended processes are actually being elicited, for example through eye-tracking and stimulated recall studies (e.g. Brunfaut and McCray, 2015; Holzknecht et al., 2017).

O'Sullivan and Weir (2011) and O'Sullivan (2015, 2016) have further modified the model, suggesting that these evidential categories can be distilled into three core areas, *the test taker*, *the test system*, and *the scoring system*. In this configuration, criterion-related validation evidence is located within the scoring system. While O'Sullivan and Weir (2011) emphasise that they consider impact to be relevant to all parts of the model and all stages of test development from *a priori* to *a posteriori* data collection, within the modifications suggested by O'Sullivan (2015, 2016), impact, values, and consequences would also be located in the scoring system, which includes evidence of the value of the scores provided by the test and the decisions made with them.

3.3 Study design

3.3.1 Overview

This study in itself constitutes a form of criterion-related validity evidence for the two tests, both in terms of how the tests compare statistically to each other, but also by triangulating the relationship between each test and the CEFR. As noted, in the original model, this type of evidence was a separate category, but has been subsumed under *the scoring system* in O'Sullivan's later adaptions. While the study in an overall sense can be posited to be a form of criterion-related validity evidence, criterial features specified under the various categories of evidence in the model were drawn on to drive the design, collection, and evaluation of evidence for the study.

As noted in Section 3.2, Davidson and Bachman (1990) made several recommendations for adaptions of the model employed in the Cambridge-TOEFL Comparability study. One caution they emphasised was the representativeness of the sample, particularly when the two tests to be compared might be designed and used in very different national and cultural contexts. The two tests in this study can be considered to have been designed in quite distinct circumstances, one an international test intended for use in various countries, and one designed and intended for use solely within one national context. However, for the purposes of this study, the focus was squarely on the use of both tests not only within one country, Vietnam, but also within one specific context of use within Vietnam, namely for graduation purposes in higher education. For this reason, the sample of test takers would be selected from test takers in that context of use, with a common-test taker design in which the same sample of university students would take both tests.

The claim of a link by both tests to the CEFR also provides an important reference point for this study, as it did in the comparison of Aptis and GEPT in the context of Taiwan (Wu et al., 2016). This common external proficiency framework provides a reference point for triangulating the scoring information provided by both tests, and comparing how both tests allocate students to the different levels of this common framework. Turning this around, if the tests are shown to provide a relatively stable and coherent determination of test takers' proficiency in terms of the CEFR, this would provide some evidence to support the claim of a link by the two tests, and suggest that a common interpretation of the CEFR levels can be obtained by two tests designed in quite different circumstances, but both taking the CEFR into account from the design state.

3.3.2 Construct definition: contextual and cognitive aspects of validity evidence

Contextual and cognitive features of the tasks in each test, fitting within *the test system* in O'Sullivan's adaptation of the model, were to be established through a comprehensive content analysis. To facilitate this analysis, an evaluation template was constructed based on the test task specifications used for Aptis. As noted above, the Aptis task specs were designed using the socio-cognitive model, and have also drawn on the taxonomies of criterial features listed in the validation studies using the socio-cognitive model noted above, such as Wu (2014), as well as the content analysis grids developed by Alderson et al. (2006). The aim was to turn the Aptis specifications into evaluative categories, mostly with a fixed number of alternatives similar to the analysis grids developed by Alderson et al. (2006), which content analysts would select from when evaluating the tests. For example, when evaluating reading tasks, in addition to contextual features, such as word count, topic, level of abstractness of the text etc., the content analysts are also asked to select the highest level of cognitive processing required to complete the items attached to the reading tasks, selecting from the six levels in the Khalifa and Weir (2009) model. The evaluation template, as with the task specifications, were divided into three substantive areas: *features of the task* (an overall evaluation of the input task and any items attached); *features of the input text*; and *features of the response* (targeting the actual items which test takers are required to complete). The template was put into an Excel spreadsheet.

Two content analysis teams of five researchers each were formed, one at the Language Testing Research Group at the University of Innsbruck (LTRGI), and another consisting of researchers from ULIS. Both teams received a day of training before carrying out an analysis of one complete form of each test. Each team was encouraged to first carry out analysis individually, then to discuss their individual results and arrive at a consensus evaluation for all tasks. This process follows that recommended by Alderson et al. (2006) when reflecting on the development of the analysis grids for reading and listening tests included in the Manual:

> *The best way of reaching agreement on the description of texts and items clearly was for analysts initially to attempt their individual analyses of the texts and items, but then to convene to discuss the individual analyses, identify sources of disagreement, and resolve differences before deciding on the definitive analysis of the texts and items* (Alderson et al., 2006:18).

The purpose of having two teams was to investigate whether two sets of researchers, in very different contexts, would derive similar interpretations of the test tasks. In particular, a key feature of the comparability study was to be how similar or different the tests were in interpretation of the CEFR. The evaluation template included an expert judgment evaluation of the CEFR level of each task, and it was thought that including expert judgments from researchers in Europe experienced in using the CEFR, as well as expert judges in the context of Vietnam, would provide interesting insights, and would also help to answer questions about the appropriateness of using the CEFR in test development outside of Europe. (Note: The issue of the appropriateness of using the CEFR outside Europe in contexts such as Vietnam is beyond the scope of this study. Both tests have established alignment with the CEFR, and this was taken as the starting point for comparing the way the tests classify students, and the way experts would judge the CEFR levels of the tests. However, Dunlea (2016a) has addressed this issue extensively in the context of using the CEFR in Japan. He has further (2016b, 2017) suggested that adapting the CEFR to local context outside Europe poses similar challenges to applying to specific contexts inside Europe, which is consistent with the original emphasis on the need for local adaption expressed in North, Martyniuk, and Pantheir (2010).)

3.3.3 The scoring system

Evidence related to the scoring system was to be collected from a number of perspectives. Descriptive statistics would be collected for both tests. However, although this would provide insights into the test performance in relation to the specific sample of test takers in this study, such descriptive statistics for separate test scores are not so informative for understanding how the different tests are classifying test takers in terms of commonly defined proficiency levels. To enable such a comparison, as both tests are linked to the CEFR and utilise the CEFR in test reporting, a comparison of the CEFR level allocation for the same students on the two tests would be carried out. To establish whether the two tests are targeting a similar construct in broad statistical terms, Principal Components Analysis was carried out using SPSS Statistics 21. More complex methods exist for examining whether the two tests are targeting a common latent variable, or variables, such as confirmatory factor analysis. However, this study was exploratory in nature, and more concerned with collating a broad range of evidence from a mixed methods design, with each piece of evidence, qualitative and quantitative, contributing to an overall picture of the similarities and differences for the two tests. For the purposes of this study, then, and also taking account of practicality and efficiency, PCA was considered sufficient to contribute some support for whether the two tests were indeed targeting a similar ability, which other aspect of the study design, such as the construct definition, might elucidate as being language ability (or not).

3.3.4 Values and consequences

As noted in Section 3.2, consequences and impact were originally presented as a separate category in Weir (2005), but are included within *the scoring system* in O'Sullivan's later adaptions, under *the value of decisions*. This study is not intended as a comprehensive validity argument for either test, but rather to contribute to such an argument for the use of the tests in higher education in Vietnam. For both tests, evidence of the impact, both intended and unintended, should be an important part of a comprehensive validity argument for such uses and will need to be collected. While it was beyond the scope of the study to investigate the impact of the two tests in terms of the wider consequences of decisions made using the tests, or in terms of washback on learning and study habits, some evidence of the values and impressions of the tests by test takers was collected through questionnaires. The intention was to collect evidence which would indicate whether test takers felt the tests were relevant to their needs, gave them a fair opportunity to demonstrate their proficiency, and also to investigate preferences towards either of the different modes of delivery (Aptis is computer based, whereas VSTEP is a pen-and-paper test with a face-to-face interview test for speaking). Importantly, for the purposes of investigating the attitudes of local students towards using an externally developed computer-based test such as Aptis, the questionnaire also asked questions about their computer familiarity.

3.4 The pilot study

Researchers from both the Assessment Research Group and ULIS first held preliminary discussions and reviewed the various documentary validity evidence available for the tests, before considering whether to proceed to actual data collection. Once it was determined that a comparability study would be both feasible and derive possibly interesting insights, a small-scale pilot study was planned. The pilot was to be carried out in one university, Vietnam National University, Hanoi. This would maximise practicality and efficient use of resources, but would obviously limit the interpretability and generalisability of results to the wider context of higher education outside this one institution. The pilot study was intended as way of trialing the procedures and data collection and analysis plans, as well as deriving greater confidence through comparison of the statistical performance of the tests as to whether it was worth proceeding to a larger scale study.

The pilot aimed to recruit 150 students who were registered to take the VSTEP test already. For these students, the use of the VSTEP was a real test event and the results would potentially be used as proof of proficiency for graduation purposes. The ULIS team invited students to participate in the study. Students were offered the opportunity to take the Aptis test using computer facilities at the university the day after they took the live administration of the VSTEP. While a counterbalanced design, with some students taking VSTEP first and others taking Aptis first would be a more robust data collection method, it was not possible. Utilising students already planning to take the VSTEP was an important way of maximising practicality and available resources for the pilot, which was essentially a feasibility study. At the same time, as the two tests were to be taken on different days, it was considered that fatigue would not be an important factor, and any practice effects of taking the VSTEP first would be ameliorated by the break in time. The span between the tests was not long enough to risk any major change in proficiency of the students, either in terms of gain or loss. Trained British Council exams staff supported the administration of the computer-based Aptis tests. One major possible limitation is the possible difference in motivation for the students, considering that VSTEP was a live test. This was an unavoidable aspect of the data collection. However, it was felt that it would be somewhat ameliorated by the fact that students taking the test were volunteers and were interested in learning their proficiency from a different instrument in the form of the Aptis test. (The corollary of this, of course, is that students in the study may in fact be more highly motivated generally than a randomly sampled group more representative of a typical cohort of students)

The students who agreed to take part would also be administered a questionnaire on their computer familiarity and attitudes towards the two tests. The questionnaire was a revised form of the questionnaire used in Wu et al. (2016) and was translated into Vietnamese by the ULIS team.

The content analysis was carried out at the pilot stage of the tests, as evaluating the similarity in terms of construct definition between the two tests was also considered an important part of determining the feasibility and appropriateness of continuing on to a larger-scale main study.

3.5 The main study

The purpose of the main study was to enhance the generalisability of the results to the wider higher education context in Vietnam. As such, it was decided to target three universities, one in Hanoi, in the north, one in Da Nang in the center, and one in Ho Chi Minh City in the south of the country. A sample of 150 students from each university would again be administered both tests and the same questionnaire from the pilot. For the university in Hanoi, ULIS at VNU was again used, and similarly to the pilot, students taking a live administration of the VSTEP were invited to take the Aptis test on the following day. Taking the tests on separate days was considered the best balance of practicality, reducing the complexity of administration for the ULIS team, as the Aptis test requires the use of computer facilities. For the remaining two universities, students for both VSTEP and Aptis voluntarily registered for the tests. The results of these tests were recognised by both host universities for their graduation. The students could choose either of the test results to gain proof of English proficiency standard. Therefore, the students were all highly motivated to take the tests. As with the Hanoi administration, VSTEP was administered on the first day, and Aptis on the second. All participating students were administered the questionnaire from the pilot study.

4. RESULTS

4.1 The pilot study

4.1.1 Scoring: comparison of CEFR level classifications

Valid test data from a total of 130 test takers who had taken both VSTEP and Aptis, was collected in the pilot study. Tables 1 and 2 present the descriptive statistics for Aptis and VSTEP respectively. Skewness and Kurtosis figures indicate that the score distributions for both tests are generally normally distributed according to rules of thumb (Bachman, 2004).

Table 1: Descriptive statistics for Aptis reported as Aptis scale scores (0–50)

	N	Min	Max	Mean	Std. dev	Skewness	Kurtosis
GV score	130	5.0	48.0	33.462	10.4749	-.595	-.711
Listening score	130	10.0	46.0	27.769	8.3104	.022	-.718
Reading score	130	8.0	50.0	32.708	10.1336	-.491	-.520
Speaking score	130	5.0	45.0	31.369	7.9164	-.838	.163
Writing score	130	12.0	48.0	38.215	7.0359	-1.072	1.227
Total score[†]	**130**	**67**	**185**	**130.06**	**29.251**	**-.474**	**-.547**

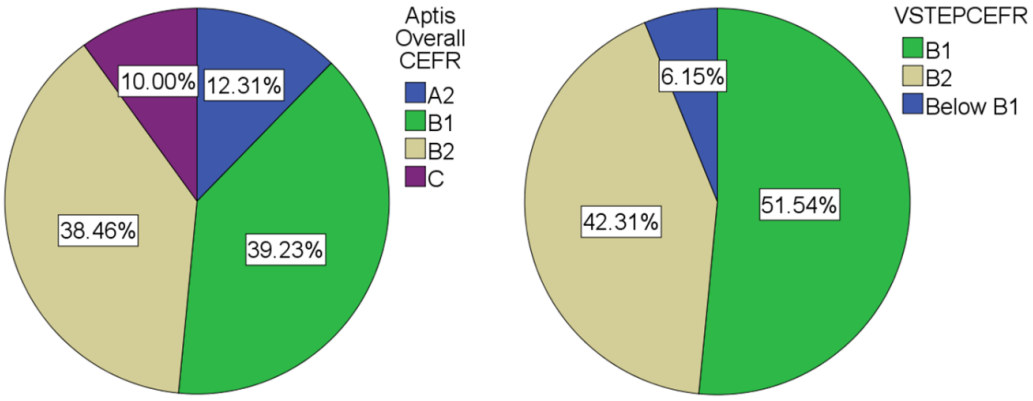
[†]For Aptis the Total Score is the sum of the 4 skills components, R+L+W+S, reported on a scale of 0-200. The Core component scale score is not included in the total score, and no CEFR level is provided for the Core component. At the same time, the Core component plays important and in many ways innovative role in the final determination of CEFR levels for the four skills, helping to take standard error of measurement (SEM) into account in borderline decisions. This is explained more fully in the *Technical Manual* (O'Sullivan and Dunlea, 2015).

Table 2: Descriptive statistics for VSTEP reported as band scores (0–10)

	N	Min	Max	Mean	Std. dev	Skewness	Kurtosis
Listening score	130	2.5	7.5	5.377	1.082	-0.264	-0.233
Reading score	130	1.5	10	6.304	1.7449	-0.531	-0.138
Speaking score	130	2.5	8.5	5.938	1.4561	-0.238	-0.931
Writing score	130	3	7.5	5.442	1.1026	-0.195	-0.763
Total score	**130**	**2.5**	**7.5**	**5.377**	**1.082**	**-0.264**	**-0.233**

As noted in Section 3.3, of particular interest in this comparability study, given both tests' use of the CEFR to inform score reporting, is the similarity or difference in the allocation of students to CEFR levels. Figures 1 and 2 provide a breakdown of the overall CEFR classification. Note that Aptis reports CEFR levels A0, A1, A2, B1, B2 and C (see O'Sullivan and Dunlea, 2015, for a detailed description of the reporting system). For VSTEP, performance below B1 is reported as below B1, with no finer distinctions. As can be seen in Figure 1, no students were allocated to A0 or A1 by Aptis. The allocation of students is broadly similar. However, Aptis appears to be stretching the students in both directions, with just over 12% allocated to A2, compared to 6.15% being placed below B1 by VSTEP. At the same time, Aptis allocated 10% of students to level C, whereas no students achieved above B2 on VSTEP.

Figure 1: Overall CEFR classification for Aptis and VSTEP



To investigate this further, Table 3 allows us to identify how the two tests allocated the same students to similar or different CEFR levels. The yellow cells indicate the number of students placed in the same levels by both tests, the green cells indicate the number of students placed into an adjacent level by one of the tests (for example 15 students have been classified as B2 by Aptis that were classified as B1 by VSTEP). Exact agreement reaches 62%. When students placed into adjacent levels are included, the agreement is 100%. In other words, although some differences in level classification occurred, no differences exceeded more than one level, with all discrepancies falling in adjacent levels.

Table 3: Comparison of CEFR classification agreement in the pilot study

		VSTEP CEFR			
		B1	B2	Under B1	Total
Aptis Overall CEFR	B1	41	7	3	51
	B2	15	35	0	50
	C	0	13	0	13
	Under B1	11	0	5	16
	Total	67	55	8	130

Key: The yellow cells indicate the number of students placed in the same levels by both tests, the green cells indicate the number of students placed into an adjacent level by one of the tests

A comparability study provides the opportunity to examine not just similarities, but also differences. Differences not just in test performance but also in terms of other features, such as how feedback is structured and its relationship to the decisions made can be equally instructive. In terms of feedback, we have already noted above that Aptis provides CEFR profiles for the separate skills components. Providing a variety of formats for reporting results was considered an important design feature from the beginning for Aptis, because of the potentially diverse nature of the global test user base, and their equally varied uses for the test (O'Sullivan and Dunlea, 2015). In the case of Aptis, although an overall CEFR level is provided for test takers, from the outset, the test developers have encouraged test users to make use of the skills profiles, as the overall level can subsume important individual differences across the skills profiles. Appendix A shows the CEFR level distributions for Listening, Reading, Speaking and Writing. For the distributions in Appendix A, it can be seen that, at least in terms of the Aptis test, this particular cohort of students appeared stronger on Reading and Writing.

For Writing, B2 accounts for the largest number of students, and both B2 and B1 receive approximately the same number of classifications. For Listening and Speaking, the largest number of students were clearly placed at B1. This is partially reflected in the band score distributions for VSTEP, with Reading having a higher mean band score than Listening or Speaking. However, for VSTEP, the Writing component also posed a significant challenge, at least for this cohort of students, a noticeable difference to Aptis.

4.1.2 Scoring: Principal Components Analysis

A Principal Components Analysis was carried out with nine variables—the nine separate test component scores for the 130 students who took both Aptis and VSTEP (five for Aptis, GV, R, L, W, and four for VSTEP, R, L, W, S). The analysis was carried out with SPSS Statistics 21. The sampling adequacy was confirmed using the Kaiser-Meyer-Olkin measure (KMO= .93). Bartlett's test of sphericity was highly significant ($p < .001$), indicating that PCA would be appropriate given the correlations between variables. The output from SPSS is presented in Appendix B. Using both Kaiser's criterion of eigenvalues greater than 1, and examination of the point of inflection in the scree plot (the scree plot is reproduced in Appendix B), the PCA provided a single-component solution, with factor loadings high for all nine variables. The results give confidence that the tests are measuring a common construct in statistical terms. It needs to be stressed, however, that the analysis does not tell us what that construct is. To inform our understanding of the tests, we need to evaluate information from the full range of evidence collected, including the content evaluation through expert judgment.

4.1.3 Questionnaire data

The results of the questionnaire are presented in Appendix C. In general, the perceptions of both tests by the university students in this study are generally positive. Most students felt the content of the tests would provide an appropriate measure of their English ability, and that the topics of both tests were relevant to their lives. The latter question is particularly important from the perspective of investigating the use of Aptis, an international test, in this local context. However, given that both tests target adult learners and intermediate to high levels of the CEFR for general English proficiency, it might be reasonable to expect that they are both tapping into a broad range of topics likely to be encountered by language users using English as a common communication tool, for example, to extract information from the Internet and international English-language media.

Noticeably larger numbers of test takers for VSTEP disagreed with the suggestion that the speech rate and accents in the Listening test were appropriate, a trend not replicated for Aptis (46% and 40% disagreement respectively for VSTEP compared to 10% and 16% for Aptis). This may be something worth pursuing by the VSTEP developers and will be touched on in Section 5. Interestingly, however, the speech rate and pronunciation of the face-to-face examiners in the VSTEP speaking test were overwhelmingly considered to be appropriate and easy to understand (94% agreeing that speech rate was appropriate and 95% agreeing that pronunciation was clear).

Although some difference in motivation and familiarity was anticipated for the two tests, in fact 56% of students said they were not familiar with the format of VSTEP; although students are aware of the test and the role it plays for them, as 74% of students said the influence of the VSTEP results on their life/work would be “quite” or “very” important. The high percentage expressing unfamiliarity might be explained by the fact that VSTEP was officially introduced just over one year before the pilot, and most of the test takers were taking the test for the first time. As it was anticipated that students may not be familiar with either Aptis or the computer-based test delivery, students were advised to visit the Aptis website and take the sample test provided to familiarise themselves with the test interface. However, 59% said they did not take the practice test. This did not seem to impede their interaction with the test, as the majority of test takers for all components agreed that the instructions for the test were clear and the interface was user-friendly (as they also did for VSTEP). Nonetheless, it would be useful for the Aptis testing team to investigate if finding and accessing the practice tests online is accessible to students attempting it for the first time, particularly if they do not have high levels of English.

For Aptis, it was important to investigate student attitudes to the delivery mode, particularly using the computer for tests of speaking and writing. Figures 2 and 3 indicate variable preferences among the students to delivery mode depending on the skill being tested. Figure 2 shows that the majority of students prefer writing on computer, as in Aptis, but prefer face-to-face speaking with a live examiner, as in VSTEP. Four questions, Q17–Q20, aimed to collect information on computer familiarity. In all four cases, the overwhelming majority suggested that: they used computers often; often read on computers; typed in English; and used computers to listen to music and read news. Interestingly, however, 75% of students said they had never taken a computer-based *test* before taking Aptis. When asked if they would prefer to take each test in an alternative mode (i.e. VSTEP by computer or Aptis by pen-and-paper), for VSTEP, a majority of students expressed a desire to take the VSTEP by computer if possible. The trend was consistent across all four skills, although the difference for Speaking was closer, with 54% to 40% expressing a desire to take the test via computer if possible. For Aptis, the reverse was true, with the majority of students answering “no” to the question would they like to take Aptis as a pen-and-paper test. Students overall, then seem to expressing openness, and indeed interest in computer-based test delivery despite their relative lack of familiarity with computer-based tests.

Figure 2: 'I prefer computer-based writing tests to paper-and-pen-based writing tests'

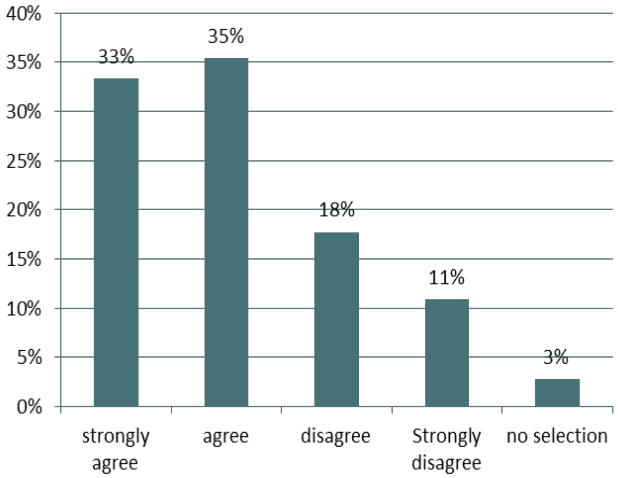
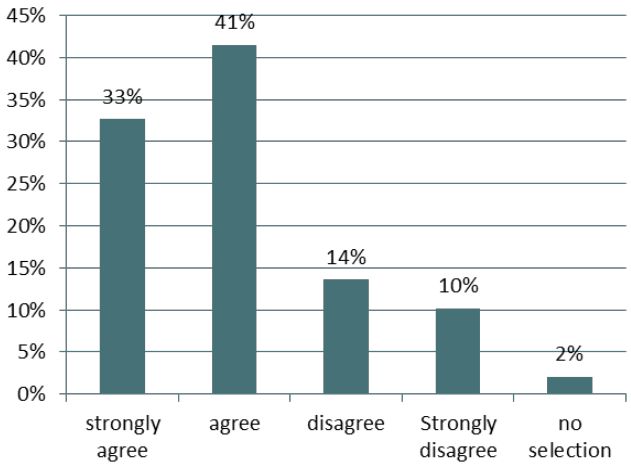


Figure 3: 'I prefer face-to-face speaking tests to machine-recorded speaking tests'



4.2 The main study

4.2.1 Scoring: comparison of CEFR level classifications

Tables 4 and 5 present descriptive statistics for the two tests from the main study. A total of 384 students across three universities took part in the main study. As noted in Section 3.5, these institutions were located in three geographically distinct areas: one in the north, one in the center, and one in the south of the country. The breakdown of sub-samples across the three institutions was: 133 from Hanoi; 118 from Da Nang; and 133 from Ho Chi Minh City. Skewness and kurtosis results, as with the pilot, indicate that on each of the components for each test, the main study population is normally distributed.

Table 4: Descriptive statistics for Aptis scale scores (0–50, 0–200 for total score)

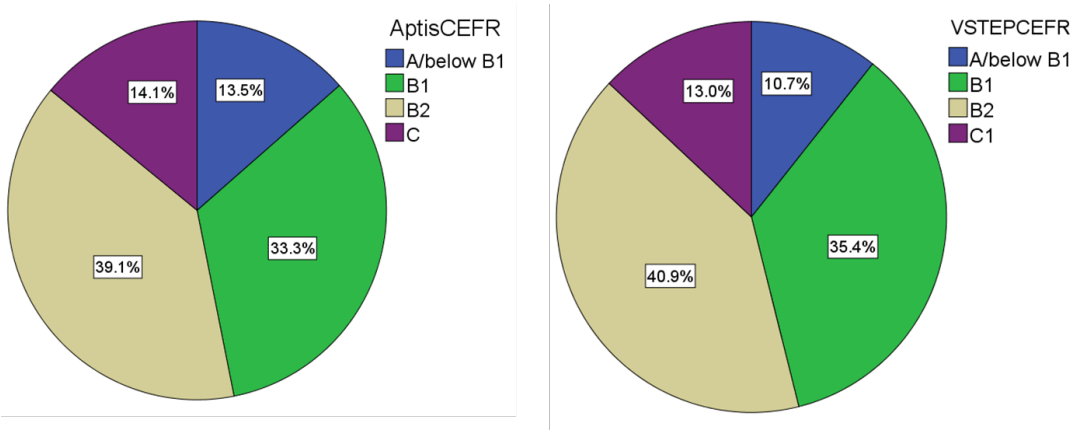
	N	Min	Max	Mean	Std. dev	Skewness	Kurtosis
GV score	384	8	49	33.362	9.2973	-0.418	-0.613
Listening score	384	8	48	28.792	9.1279	0.02	-0.696
Reading score	384	4	50	33.599	10.7619	-0.468	-0.459
Speaking score	384	0	48	31.974	9.6202	-1.001	0.97
Writing score	384	4	50	37.865	8.4286	-1.53	2.417
Total score	**384**	**42**	**194**	**132.229**	**33.6772**	**-0.622**	**-0.098**

Table 5: Descriptive statistics for VSTEP band scores (0–10)

	N	Min	Max	Mean	Std. dev	Skewness	Kurtosis
Listening score	384	1	9.5	6.09	1.9991	0.144	-1.129
Reading score	384	2.5	10	6.585	1.8837	-0.257	-0.877
Speaking score	384	0	10	6.009	2.083	-0.16	-0.823
Writing score	384	0	9	4.811	1.7067	-0.136	-0.253
Overall score	**384**	**1.5**	**9.5**	**5.936**	**1.7476**	**0.025**	**-0.942**

In terms of classification into overall CEFR levels, the breakdown in the main study is much closer than in the pilot. For the main study, Aptis results under B1 (A2, A1, or A0) were subsumed into one category, below B1, to facilitate comparison with VSTEP. As can be seen from Figure 4, the level allocations are much closer, with both tests allocating the largest number of students to B2 (39.1% for Aptis and 40.9% for VSTEP), with very similar numbers being put into B1 (33.3% for Aptis and 35.4% for VSTEP). Both tests also allocated very similar percentages of students to both below B1 and to the C1 category.

Figure 4: Overall CEFR classifications for Aptis and VSTEP



Taking a closer look at classification agreement in Table 6, we can see that there is 66% exact agreement on CEFR level classification for students by the two tests. Adjacent agreement – in which a student has been placed into an adjacent level by one of the tests – is 33%, giving a combined total of 99% for exact and adjacent agreement. As with the pilot, this indicates fairly broad agreement in level classification by the two tests. A very small number of students were placed two or more levels differently, with two students being placed at B2 by VSTEP that were placed below B1 by Aptis, and three students placed at B2 or C by Aptis that were classified as below B1 by VSTEP. Although these cases account for only 1.3% of students, they will require further investigation. It is possible that students experienced some problem in one of the tests, with technical difficulties being one possibility for the computer-based Aptis.

Table 6: Comparison of CEFR classification decisions in main study

AptisCEFR * VSTEPCEFR Crosstabulation

Count

		VSTEP				
		below B1	B1	B2	C1	Total
Aptis	**A/below B1**	26	24	2	0	52
	B1	12	90	24	2	128
	B2	2	22	107	19	150
	C	1	0	24	29	54
Total		41	136	157	50	384

Key: The yellow cells indicate the number of students placed in the same levels by both tests, the green cells indicate the number of students placed into an adjacent level by one of the tests

Figures 5 and 6 provide an overview of the three university sub-samples within the main study population. Figure 5 shows box plots of the total score performance (for Aptis, this is a sum of the four skills scale scores on a scale of 0–200, for VSTEP this is an overall band score of 0–10). The box plots indicate that there are no major differences between the samples, with a generally similar pattern for both Aptis and VSTEP across the three universities. For both tests, Ho Chi Minh City has the highest median score but also the greatest variability in terms of score range. Da Nang has the second highest median score and the most restricted range. Despite some differences, however, the three universities show a great deal of overlap.

Figure 6 shows the breakdown of overall CEFR level allocation, with some slightly more pronounced differences. The Ho Chi Minh City sample shows a much larger percentage of students allocated to the C level by Aptis. This trend is not maintained in VSTEP, where the number of students allocated to C1 is greatest for Hanoi, although the results are quite close for Ho Chi Minh City and Hanoi. The restricted range for Da Nang in terms of total scores is reflected in the CEFR profiles. For both tests, approximately 90% of the Da Nang sample is classified as B1 or B2, with much smaller percentages than the other universities falling into Below B1 or C1.

The broadly similar nature of the three university groups in terms of ability is perhaps not surprising and possibly a feature of the nature of the sample which was a convenience sample. All three participating universities are among high-performing institutions. While this gives us confidence that the three sub-samples within the main study population are broadly comparable, it does pose some limitations on generalisability of results, which will be discussed in Section 5.

Appendix D provides the profile of CEFR levels for each component of Aptis. The difference in proficiency allocation is much less pronounced for Reading and Listening, with both components allocating the largest number of students to B1, but with B2 being the second largest category, only slightly lower than the percentages for B1. Together B1 and B2 account for around 60% of students for both Listening and Reading. Speaking is by far the most challenging, with 60% of students allocated to B1, while this cohort on this version of Aptis, show a similar pattern to the pilot, performing most strongly on Writing, with almost 50% of students reaching B2.

Figure 5: Box plots for Aptis and VSTEP total scores

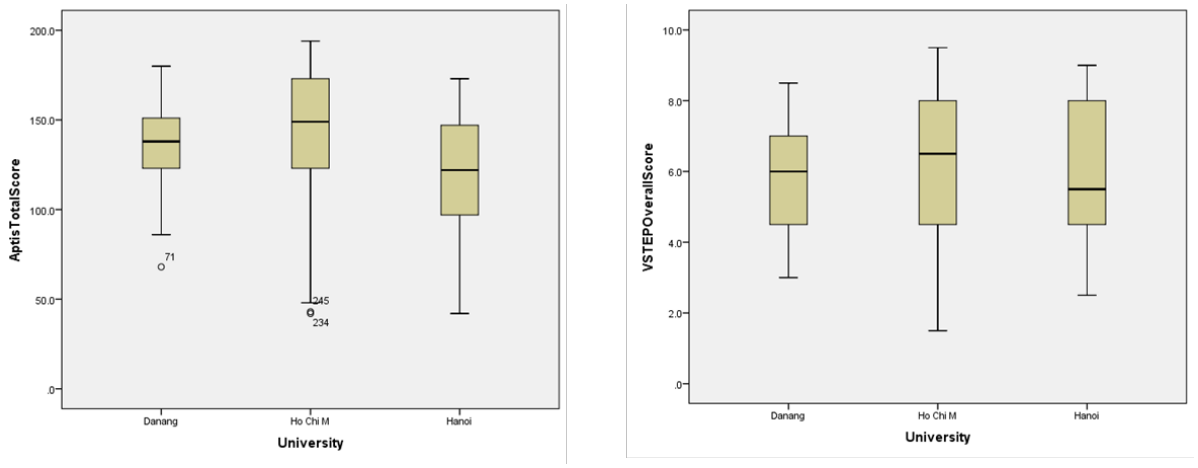
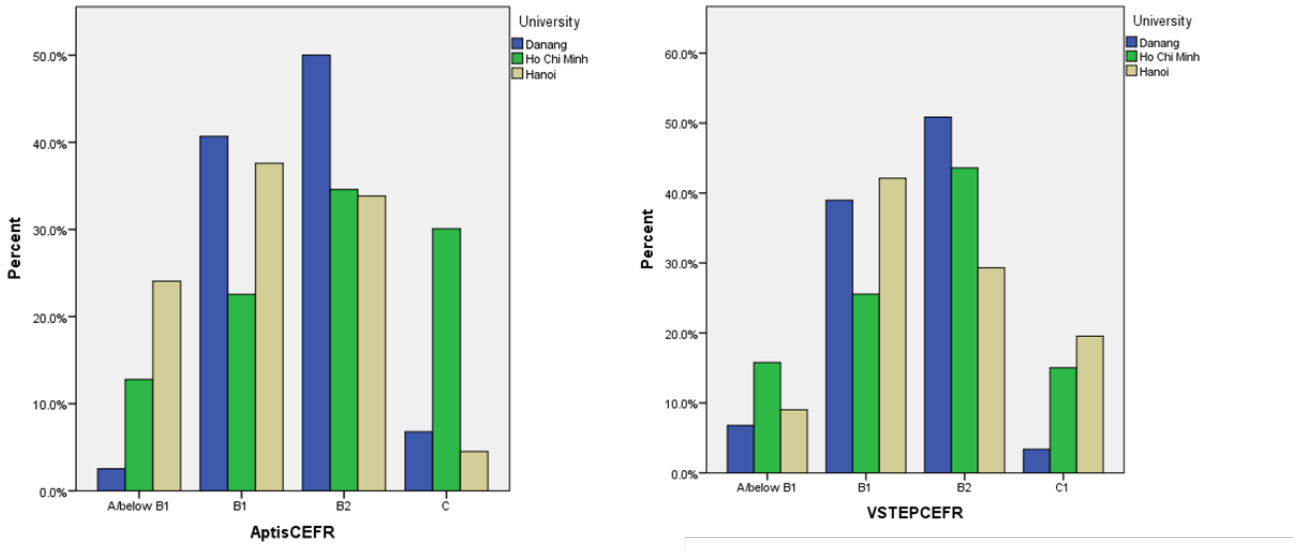


Figure 6: Breakdown of overall CEFR classification by university



4.2.2 Scoring: Principal Components Analysis

To investigate whether the Aptis and VSTEP testing instruments would behave in the same way as in the pilot with this more diverse sample, spread across there institutions in different geographic locations, a Principal Components Analysis (PCA) was carried out. The PCA used the same nine variables as in the pilot—the nine separate test component scores—this time for the 384 students who took both Aptis and VSTEP in the main study. The output is presented in Appendix E. The analysis was carried out with SPSS Statistics 21. The sampling adequacy was confirmed using the Kaiser-Meyer-Olkin measure (KMO= .941). Bartlett's test of sphericity was highly significant ($p < .001$), indicating that PCA would be appropriate given the correlations between variables. Using the same criteria as in the pilot, Kaiser's criterion of eigenvalues greater than 1 and examination of the point of inflection in the scree plot, the PCA in the main study once again provided a single component solution, with factor loadings high for all nine variables. The results give confidence that, with this more generalisable sample, the tests are still measuring a common construct in statistical terms.

4.2.3 Questionnaire data

In general, the trends observed in the pilot questionnaire data are replicated in the main study. Students were generally positive in their impressions of both tests, considering that the results of the tests would reflect their ability, and that the topics in both tests were relevant to them. As noted in the description of the pilot questionnaire results, it is particularly important for the Aptis test developers to investigate whether the content of this international test used in globally varied contexts is accessible and relevant to test takers in this local context of higher education in Vietnam. The questionnaire results from the main study give some support to this suggestion.

Regarding the speech rate and accents used in the listening test, approximately 30% of students disagreed with the suggestion that the rate and accents for VSTEP are appropriate — slightly lower than in the pilot, but still noticeable numbers of students. Once again, students did find not this to be the case for the speech rate and pronunciation for the speaking examiners, with the overwhelming majority considering the examiners' speech rate appropriate and the pronunciation clear.

A larger percentage of students than in the pilot said that they were not familiar with the format of the VSTEP test (70% in the main study compared to 56% in the pilot), reflecting the fact that the samples outside of Hanoi were not taking a live VSTEP test, as were the students in Hanoi for the main study and the pilot. Nonetheless, a majority of students, 70%, still considered the VSTEP to have important consequences for their life and work.

For Aptis, we once again see that the majority of students, 72%, did not take the freely available online practice test before taking the Aptis test. As with the pilot, although this did not seem to impede students' interaction with the test, with the majority considering the instructions to be clear and the interface accessible, it would be useful for the test developers to consider how to make access to the practice tests more accessible to test takers.

In terms of computer familiarity, the overwhelming majority of students answered positively for the four questions targeting this aspect. Over 80% of students said they use computers often, often read on computers, type in English and use computers or digital devices to listen to music and read the news. At the same time, as with the pilot, 66% of students said they had not taken a computer-delivered test before taking the Aptis test.

A majority of students expressed a preference for computer-based writing tests over pen-and-paper writing tests, but as with the pilot, a majority preferred face-to-face speaking tests to the semi-direct, pre-recorded delivery in the Aptis test. When asked if they would prefer to take each test in alternative delivery mode, the same trend was evident as in the pilot, with the majority of students responding "yes" for all components of the VSTEP. Though once again, the difference was much less for Speaking. For Aptis, the majority of students continued to express a preference for taking Aptis as a computer-based test over a pen-and-paper delivery mode.

The replication of trends seen in the pilot questionnaire data with the three different universities in the main study, as with the test performance breakdown shown in Figures 5 and 6, points to broad similarities in the sub-samples making up the main study population.

4.3 Construct definition

The construct definition aspect of the project generated a great deal of data. As noted in the methodology section, two teams of five researchers each were trained, one based in Europe and one at ULIS in Vietnam. The construct definition took place at the pilot stage, and for reasons of practicality, was only carried out once; this will be discussed in more detail in Section 5. Each content analyst used a standardised evaluation template based on the literature on the socio-cognitive model, and drawn specifically from the criterial contextual and cognitive parameters contained in the Aptis test specifications. A separate analysis template in Excel format was used for each test component for Aptis and for VSTEP. This generated a total of 10 content evaluations for each component for each test (90 in total, with five components for Aptis and four for VSTEP). Following individual analysis, each team carried out a group discussion in which they arrived at a consensus evaluation for each component, yielding a further 18 evaluation forms, one for each research team for the five Aptis components and four VSTEP components. In addition, a separate automated analysis was carried out to generate readability estimates and vocabulary level profiles for the input texts used in the reading and listening components of each test.

A detailed analysis of this data is beyond the scope of this report, and will be distilled into a separate report focusing on the results, and the process, of the construct evaluation, reflecting its importance within the study design and the socio-cognitive approach that underpins it. In the context of this report, we will summarise main trends, and focus in detail on one specific example, comparing the consensus analysis of the final long reading task in both tests to demonstrate the kind of information the evaluation templates can yield, and the areas of similarity and difference they can help identify, both between the tests and also between the two analysis teams. The separate full report on the construct definition stage will look at not just the consensus forms, but also investigate differences in analyses across individuals, looking for individual trends, as well as trends that might be associated with a particular team.

Appendix G presents an example of a full Excel analysis template, formatted for the Aptis long reading task. The template can be adjusted to accommodate different tasks types with different numbers of items. The criterial contextual and cognitive parameters vary depending on the skill focus, reading, listening, writing, or speaking. The template is split into three major areas of analysis, *features of the task, features of the input* text and *features of the response*. For the majority of categories, the options for selection come from fixed sets of options derived from the literature on uses of the socio-cognitive model, and distilled specifically from the Aptis test specifications.

Appendices H to K focus specifically on the consensus evaluations from both the LTRGI team and the ULIS team for *features of the task* and *features of the response* for the final long reading task for both Aptis and VSTEP. *Features of the input text* includes many of the automated analyses mentioned above, such as vocabulary level, and will not be examined in detail here as the results for this section will be dealt with in more depth in the independent report on construct definition to be completed separately. Looking at the features of the task for both Aptis and VSTEP, we see broad agreement between both teams regarding key features. Firstly, based on expert judgment, both teams have judged the holistic, overall task level to be the same, B2 for Aptis and C1 for VSTEP. This is generally consistent with the test design. For Aptis, each of the four reading tasks is intended to target a specific CEFR level, A1, A2, B1, or B2 (see the Aptis test specifications in O'Sullivan and Dunlea, 2015). VSTEP extends the levels targeted to C1, although as we shall see when discussing features of the response, there are also important design distinctions between the two reading tasks that the evaluation template can help us elucidate.

The two teams have also agreed on the key cognitive processing demands of the tasks — global expeditious reading and building a mental model for Aptis, and global careful reading and text level representation for VSTEP. These categories are based on the Khalifa and Weir (2009) cognitive processing model of reading which has been adapted and incorporated into the Aptis test specifications (see O'Sullivan and Dunlea, 2015, and Brunfaut and McCray, 2015, for more detailed discussions of this model and its validation in the context of Aptis reading tasks). As noted in the literature review on the socio-cognitive model, the incorporation of cognitive processing into test specification has been an important contribution of the model and its operationalisation in the Aptis test specifications. As such, seeing broad agreement on these key features for both tests by the two teams gives us confidence in the ability of the teams to make coherent judgments using the templates, and that the reading tasks are indeed targeting a similar kind of EFL reading ability. The consistency in CEFR level classification also indicates that, based on expert judgment by these two teams working independently, the tasks appear relevant to specific CEFR levels. In features of the task, the two teams differed only in their analysis of *content knowledge*, with the ULIS team indicating that they considered both Aptis and VSTEP reading tasks to require a higher level of background knowledge than did the LTRGI team.

Two key parameters have been included here from features of the response, the *key information* and *the operation*. For features of the response, a separate analysis is made on these key categories for every item attached to a task. For these long reading tasks, there are seven items in Aptis and 10 items attached to the VSTEP task. Alderson et al. (2006) noted that in their study, analysts were only able to achieve moderate levels of agreement for the operations targeted by specific items. The results here show that, for operation, the two teams differed completely in the analysis of Aptis items, with the LTRGI team selecting *main idea / conclusion*, whereas the ULIS team selected *gist* for all items. However, for the VSTEP task, the two teams had perfect agreement on the operation targeted by each item.

The *key information* category is a crucial part of operationalising cognitive processing as a part of test specification, and here in item classification. Key information refers to the degree of integration across a text that is required in order to complete an item successfully; where is the information required to answer successfully located in a text, within a single sentence, across sentences, or across paragraphs. This parameter operationalises the Khalifa and Weir model (2009) that has been incorporated into Aptis test specification. Within the Aptis test design, a link is made between the highest level of cognitive processing required by a task and reading appropriate to different CEFR levels. This link has a quantitative and qualitative dimension, with A1 reading targeting within-sentence comprehension at the word, phrase and sentence level for very short texts, and B2 reading targeting the integration of information and propositions across longer stretches of text, with the total length of reading texts also being much longer than tasks targeting lower CEFR levels (see O'Sullivan and Dunlea, 2015, and Brunfaut and McCray, 2015, for more details). For Aptis then, reading considered appropriate for B2 would require at least across-sentence comprehension. Both teams concurred that all items targeted this level for Aptis. While this meets minimum requirements, the test specifications for Aptis indicate that across-paragraph reading is also being targeted at this level, indicating that the expert judgment analysis of this task indicates it may not be pushing the full bounds of B2 as intended by the test developers. For the 10 items in the VSTEP task, the two teams concurred on 8 out of 10 items, differing only on items 4 and 8.

The analysis template is instructive also in helping us to understand key differences in test design. For example, as noted above, for Aptis, B2-level reading requires higher cognitive processing demands, operationalised with a minimum of across-sentence comprehension required at an item level (it is important to note, however, that these features alone do not determine the level, and that classification at B2 level would be in conjunction with other key criterial B2-level features, such as an appropriate level of text abstractness). However, although both teams classified the VSTEP reading task holistically as C1, several items were classified by both teams as targeting within-sentence processing, which would run counter to the higher-processing demands for B2 and C-level reading posited by the Aptis test design. However, reference to the test design for VSTEP makes it clear that

within the larger number of items attached to the task (10 for VSTEP compared to 7 for Aptis), several items deliberately target *vocabulary knowledge*. For Aptis Vocabulary and Grammar items are separated into a stand-alone component, the Core component. For VSTEP, such items are incorporated into the reading tasks, and as they are not targeting reading comprehension *per se*, would not be expected to require the same cognitive demands as the items targeting reading comprehension. Within the VSTEP task overall, we see a number of items targeting reading comprehension at both across-sentence and across-paragraph levels, helping to explain the holistic judgment of C1 by both teams for this task. The template is thus instructive in helping us to elucidate both differences and similarities in test design.

While the full analysis will be presented, as noted above, in a separate report, several other important trends will simply be noted here. Across the remaining reading tasks in both texts, and also across listening, writing and speaking tasks, similar trends were noted, with broad agreement on key characteristics between both teams. The teams were generally in agreement on the CEFR level targeted and, where disagreement occurred, it was generally adjacent agreement. The range of CEFR levels targeted generally concurred with the overall test design, with tasks up to B2 in all components of both tests. Aptis included more tasks at lower (A1 and A2) levels, which is once again consistent with the design of both tests. Apart from Reading, however, VSTEP did not include C1 level tasks, according to these two sets of expert judgments (apart from one item in listening classified at C1 by the LTRGI team). This may have implications for the VSTEP test developers given the aim of VSTEP to target levels B1 through to C1. However, it needs to be stated that this is only one piece of contributing evidence, and comes from expert judgment. Other evidence, including the psychometric properties of the test tasks, and a wider scope of content evaluation as recommended in Section 5, would be needed before making any decisions about the level of tasks. There was also a relatively high level of agreement on the important feature of key information for both listening and reading, although as with the reading task analysed in detail here, with some discrepancies at the individual item level.

This summary of features focusing in depth on only one reading task from each test has been used to demonstrate that the analysis template appears to provide an accessible and useful way of analysing test content, including on key cognitive parameters which have been shown to be useful for identifying and validating test tasks (e.g. Brunfaut and McCray, 2015; Holzknecht et al., 2017).

5. DISCUSSION

5.1 Limitations

There are certain limitations to the study which require noting. Firstly, although efforts were made to broaden the generalisability and interpretability of results by including three universities in different locations, the sample still has limitations in terms of representativeness for the wider higher education sector in Vietnam. The three universities included were considered high-level institutions within the context of Vietnam and were amenable to participation through existing links and collaboration with ULIS. We have already referred to the similarities in terms of both the test performance data and questionnaire data for the three samples, and this may be a reflection of shared features of these institutions. Within each of the institutions, the sample is also a convenience sample. Although efforts were made to recruit students from a range of levels and classes, in the end, as is often the case, recruitment depended on students willing to participate, meaning the sample cannot be considered a randomly selected sample within each institution. This has some implications for generalising the results to the wider higher education sector, which will include a wider range of institutions catering to students of varying levels of ability.

Another limitation refers to the number of test forms used for both test data collection and content analysis. Davidson and Bachman (1990), in their recommendations for future applications of the comparability model they suggested could be based on the CTCS, recommend that using multiple test forms would strengthen the interpretation of results. However, in this study, for the test delivery, the VSTEP was, in fact, a live administration in the pilot, and a live administration for the Hanoi group in the main study. This then limited the choices of test form for VSTEP to that used in that live administration. For Aptis, live administrations utilise a bank of fully pre-tested and pre-equated test forms which are randomly allocated to test takers, including test takers in the same room (see O'Sullivan and Dunlea, 2015, for details). These test forms are considered to be statistically comparable and interchangeable in terms of difficulty. However, to streamline the analysis and avoid any possible test form affects when comparing the CEFR classifications for the two tests with the limited number of students it would be possible to recruit, it was decided to utilise only one form of Aptis for data collection. The form used was a live Aptis test form. The same Aptis test form was used in both the pilot and main studies for data collection.

For content analysis also, as already explained above in the Methodology section, one complete test form was analysed for each test. The logistic constraints of recruiting two teams of researchers in Europe and Vietnam, training them, and then having them carry out individual analyses before coming together in their respective teams to discuss and reach consensus were considerable. It was not feasible, given the time demands on each team, to require them to carry out the analyses on more than one form for each test. While the use of single test forms does mean that care needs to be taken when generalising from the results of this study, all live test forms in both tests are built to the same specifications in terms of content. For Aptis, task specifications are extremely tight and production is overseen by a central team of specialists and quality assurance managers (see O'Sullivan and Dunlea, 2015, for details), which would somewhat ameliorate the concerns expressed by Davidson and Bachman (1990). For VSTEP, a particular feature of the current test model is that a set number of certified universities may in fact produce test forms for administration in their institution according to the same specifications produced by the original test developers at ULIS. For VSTEP, then, the implications for form comparability and generalisability from this study to other forms of VSTEP produced across institutions may be more serious. This will be taken up further in the Recommendations section below.

This study was meant to capture a broad range of different aspects of evidence to enable a comprehensive comparative analysis of the two tests, while also investigating the utility of the study design itself and instruments, such as the content analysis template. A balance, therefore, needed to be struck between the various demands of the study, including practicality, generalisability of results, breadth of evidence, and depth of analysis. For the purposes of this study, we feel an acceptable balance was obtained, but recognise the limitations that are entailed by these design decisions.

5.2 Main findings

As outlined in Section 1.4, the study had two particular dimensions, an instrumental dimension of interest particularly to the developers of each test to contribute evidence towards the validity argument to justify the use of each test within the context of higher education in Vietnam. For Aptis, comparability to a locally used testing instrument and investigation of the impressions of typical test takers in this context was important. For VSTEP, a locally developed test that claims relevance to international benchmarks of performance in terms of the CEFR as required by government mandate, investigating whether those claims would play out in relation to an international test already linked to the CEFR, would add weight to the test developer's arguments. The relationship to the CEFR claimed by both tests was seen as an important feature that would allow the triangulation of test results to facilitate a comparison of not just test performance, but also in terms of expert judgment content analysis. The second key dimension to the study was to further pilot and refine a methodology and set of instruments based on the socio-cognitive model for test development and validation. This wider goal aimed to look beyond this particular context in order to develop an explicit and coherent framework for such comparability studies, particularly for tests claiming a link to an external proficiency framework such as the CEFR.

From both of these perspectives, the study has provided important information. Looking first to the more instrumental goals of the study, the scoring information obtained has suggested that the two tests do indeed provide very similar interpretations of overall CEFR level classifications, and the Principal Components Analysis added confidence that the two tests are tapping into a similar underlying construct. The fact that these trends were generally repeated across the pilot and main studies, adds further weight to their interpretation. The assumption that the common construct of the two tests identified statistically is, in fact, related to a general EFL language proficiency relevant to adult EFL learners, and that this proficiency can be further described in terms of CEFR levels, was given support from the construct definition through the content analysis templates, with independent judgments from two separate research teams in Europe and Vietnam arriving at similar interpretations, particularly in relation to CEFR levels targeted. The questionnaire data provided some confidence that students found the tests accessible, and felt that the content was relevant to them.

At the same time, the content analysis templates provided insights into important differences in test design. Interpreting these differences, and evaluating them in terms of the appropriateness for the uses and interpretations posited for each test, will require further analysis. This would usefully include comparing the expert judgments derived here to the test specifications and intentions of the test developers, and further elucidation by the test developers as to why the particular design decisions for each test were taken and how they support the interpretations recommended for the test.
The questionnaire data also yielded interesting insights into this particular group of students, which although it might not be completely representative of the higher education sector, does span three distinct geographical locations representing major areas of Vietnam. The data revealed interesting insights into the computer familiarity — very high — of these test takers. While they conversely had little experience with computer-based tests, the questionnaire data seems to indicate that this particular cohort is very open to the option of computer-based testing (although this is less so when considering speaking).

Difference in motivation was initially considered one potential factor which could possibly lead to differential performance on VSTEP and Aptis, given the status of VSTEP as a national, standardised test. However, this does not appear to have been a major factor, with very similar performance levels and CEFR classifications across both tests indicating similar levels of effort being applied by the test takers in the study. As noted in the Methodology section, the study design gave students in the two universities outside Hanoi the choice of using either test as official results for graduation purposes, and this, by design, is likely to have given students similar motivation for both tests. At the same time, the questionnaire data, particularly for the main study, indicated that large numbers of students were not familiar with the format of either test before participation in the study, indicating no particular impact for increased familiarity for what would be considered the local test over the international test. We can cautiously conclude, then, that students approached both tests with similar levels of motivation.

In relation to the second major rationale for this study, developing a methodology for comparability studies, the socio-cognitive model has provided a coherent and comprehensive framework for guiding the selection of evidence across a fixed set of validation aspects which together contribute to an overall picture of construct validity. This is not surprising when one considers that this particular study has built on a broad body of literature employing the socio-cognitive model, particularly drawing on Wu (2014), Taylor (2014), Chan and Taylor (2015), Dunlea (2016a), and experience of the development of the Aptis (O'Sullivan 2015a, 2016, O'Sullivan and Dunlea, 2015). This particular methodology as noted in the literature review was also piloted in Wu et al. (2016), and further adapted for this study, in particular, for the content analysis templates.

The purpose of the content analysis was two-fold. Firstly, it was designed to add clarity to statistical and other comparisons investigating comparability. Without an understanding of what is being targeted in each of the tests, including the interaction between items and input texts, not just in terms of superficial textual features of the input texts, it is impossible to truly identify similarities or differences. At the same time, this study aimed to build on and refine the evaluation templates originally derived for use in Wu et al. (2016), and to develop a model for content analysis which would capture the key criterial contextual and cognitive parameters stressed in the literature on the socio-cognitive model, and which, if applied across test comparison projects, offers the potential to facilitate comparison and evaluation across contexts using consistent and well-defined categories. As already noted in the literature review, O'Sullivan (2017) has emphasised the importance of such construct definition, and the important place it holds in the socio-cognitive model, and recommended it needs to be given a much higher priority in any study attempting to link tests to an external proficiency framework. The instrument presented in this study is offered as a potentially useful tool for that purpose.

5.3 Recommendations

A number of recommendations suggest themselves from the analysis and evaluation of the results so far.

Firstly, this study was intended to prioritise collection of a broad range of evidence to facilitate a multi-faceted approach spanning both qualitative and quantitative aspects, as noted above. As such, breadth was, to some extent, prioritised over depth in analysis. However, the wealth of data derived would be amenable to further in-depth analysis without collecting any more data. For example, it would be useful to compare test performance for the three groups in the main study to wider performance data for VSTEP nationally, including those forms delivered by (and developed by) other institutions. For Aptis, it may be instructive to compare the sample in this study to characteristics of similar test user groups, i.e. university students, in other higher education contexts, as well as to compare characteristics of this sample to performance data for other Vietnamese institutions using Aptis for live administration purposes.

Further analysis of existing data could also include Rasch analysis to place test items from both tests onto a common scale of difficulty. Item difficulty on such a common scale could then be cross-referenced with the intended CEFR levels of tasks and items, the CEFR cut-offs set at the test level separately through standard-setting procedures, and the expert judgments of items and tasks obtained through this study.

Secondly, if feasible, it would be instructive to replicate the content analysis used for the construct definition stage with a wider number of test forms for both tests. Indeed, for VSTEP, it may be particularly instructive to use the analysis templates to compare test forms across VSTEP administrations, particularly forms produced by other institutions, as noted in Section 4.3.

Thirdly, it would aid interpretation and validation not just of the results of this particular study, but in the ongoing development of a validity argument for the uses and interpretations of VSTEP to increase the amount of publicly available documentation on the development and validation of the VSTEP test. As noted in the literature review, the VSTEP project has built a number of strong collaborative relationships with international researchers. However, there is little detailed published documentation in the form of technical manuals, technical reports, or test specifications aimed either at researchers or test users such as teachers and learners. Aptis has made a point of publishing detailed test specifications, something large-scale test developers are sometimes reluctant to do, as it is sometimes seen as carrying the risk of construct-irrelevant variance through encouraging test-preparation strategies. From the Aptis test development point of view, detailed public test specifications enhance teaching to the construct, not the test, and are an important part of ensuring validity.

Improved public documentation for VSTEP would facilitate better understanding of the test amongst key stakeholders and user groups. Ensuring at least some of that information is available in English will further facilitate collaboration with international researchers. While VSTEP is a national, standardised test for the context of Vietnam, a key premise of its design is relevance to external frameworks of reference, including the CEFR and CEFR-VN. Facilitating research and understanding at an international level can be a useful way of demonstrating that relevance to local stakeholders. Given that one of the key goals of the VSTEP project was to build capability in test development and research expertise, it is important that the project follows up on the important work it has delivered so far to further set high standards in terms of professional standards and accountability. Providing more thorough public documentation, from technical performance data to test specifications, will further enhance the development of language assessment literacy and professionalism in language testing in Vietnam. This is particularly true in terms of alignment with external proficiency frameworks, such as the CEFR, and the VSTEP can make a major contribution to the local context by demonstrating more clearly in public documentation how this was done according to best practice.

Linked to the above recommendation, for both tests a relatively large number of students said they were unfamiliar with the VSTEP or had not taken the available online practice test for Aptis. Both tests could possibly benefit from further investigation into how to improve accessibility to information on the tests and particularly to practice materials for familiarisation for test takers. While Aptis has actually produced a relatively large amount of documentation—a detailed technical manual and a number of technical and research reports, as well as candidate guides—much of this information is in English and technical in nature. Producing more material aimed at non-specialists, particularly in their first language, may facilitate the accessibility of the Aptis test in varied local contexts. In addition, technical performance data in the *Aptis Technical Manual* and other documents is often based on a global data set. Given the importance placed on localisation in the Aptis test development model, it would be beneficial to repeat studies such as this in local contexts with particular sets of test users.

5.4 Conclusion

The study has provided important insights into the use of Aptis and VSTEP in the context of higher education in Vietnam. It has yielded a large amount of data which provides the possibility of further in-depth analyses. It has highlighted not only similarities, but also differences, in test design and delivery. We would not expect, or want, two tests developed in different context to be identical. Understanding those differences and the implications they have for the uses and interpretations of the tests is important. Collecting evidence across a range of categories aids this deeper understanding. For example, deriving expert judgments of CEFR levels is instructive and helps aid interpretation of statistical difficulty, but digging deeper through the content analysis templates provides the possibility of building a much more comprehensive picture of the criterial features characterising CEFR levels for each test. Even two tests developed with a claim of alignment to the same levels of an external framework such as the CEFR can be expected to contain important differences. The CEFR has, in addition to the Global Scale, 54 separate illustrative scales targeting different aspects of language use and competence. No test realistically would expect to target all of those scales in test development. So even two tests targeting the same level, for example B2, and with sufficient evidence to support their claim of alignment, would be expected to demonstrate differences in content and design. Detailed test specifications underpinning item writing and quality assurance are a useful way of providing that kind of information about a test separately to test users. To further enhance validation through external review, or indeed to develop such specifications *a posteriori* where less detailed specifications may have been used for initial test development, the content analysis template used for construct definition in this study provide a potentially powerful tool.

The collaborations built through the project have engaged not just the Assessment Research Group at the British Council, the British Council in Vietnam, and the test research team at ULIS, but have also involved the Language Testing Research Group at Innsbruck University and the other institutions participating in the main study. As such, not only has the study provided the opportunity to contribute to the wider field through demonstrating the utility of the comparability study methodology suggested, it has built professional networks across these diverse groups that promise further useful and insightful research possibilities in the future.

REFERENCES

Alderson, J.C. (2005). Editorial. *Language Testing,* 22 (3), 257–226.

Alderson, J., Figueras, N., Kuijper, H., Nold, G., Takala, S. & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference: the experience of the Dutch CEFR construct project. *Language Assessment Quarterly,* 3(1), 3–30.

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

Bachman, L.F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bachman, L.F., Davidson, F., Ryan, K. & Choi, I-C. (1995). *An Investigation into the Comparability of Two Tests of English as a Foreign Language*: The Cambridge TOEFL Comparability Study, Studies in Language Testing: Vol. 1. Cambridge: Cambridge University Press.

Brunfaut, T. & McCray, G. (2015). *Looking into test-takers' cognitive processes while completing reading tasks: A mixed-method eye-tracking and stimulated recall study.* ARAGs Research Reports Online, AR-G/2015/001. London: British Council. (See: https://www.britishcouncil.org/exam/aptis/research/publications/non-technical-reports)

Carr, T. Nathan, Nguyen, T.N. Quynh, Nguyen, T.M. Huu, Nguyen, T.Q. Yen, Nguyen, T.P. Thao. (2016). *Systematic support for a communicative standardised test of proficiency in Vietnam*. Paper presented at the 4th New Directions in English Language Assessment Conference. Hanoi, Vietnam.

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

Chapelle, C. A., Enright, M. K. & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, *29*(1), 3–13.

Council of Europe. (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe. (2009). *Relating language examinations to the Common European Framework of References for Languages: Learning teaching, assessment.* Strasbourg: Language Policy Division.

Davidson, F. & Bachman, L. (1990). The Cambridge–TOEFL Comparability Study: An example of the Cross-National Comparison of Language Tests. In J. de Jong (Ed.), *Standardisation in Language Testing: AILA Review 7*. Accessed from http://www.aila.info/download/publications/review/AILA07.pdf

Davidson, F. & Fulcher, G. (2007). The Common European Framework of Reference (CEFR) and the design of language testes: A matter of effect. *Language Teaching*, 40, 231–241

Dunlea, J. (2016a). Validating a set of Japanese EFL proficiency tests: demonstrating locally designed tests meet international standards. Unpublished PhD thesis. University of Bedfordshire, Bedfordshire.

Dunlea, J. (2016b). *Tensions and synergies in standardised testing: making the numbers meaningful.* Keynote presentation at the 4th New Directions in English Language Assessment conference, Hanoi, Vietnam, October.

Dunlea J. (2017). *Setting standards: the role of assessment in implementing the CEFR in education reform.* Keynote presentation at the 8th International Conference on TESOL in Vietnam, jointly organised by the SEAMEO Regional Training Center in Vietnam and Curtin University in Australia. Ho Chi Minh City, Vietnam, August.

Dunlea, J., Nguyen, T.N. Quynh, Nguyen, T.M. Huu, Nguyen, T.Q. Yen, Thai, H.L. Thuy, Nguyen, T.P. Thao (2016). *Reporting on the pilot phase of a test comparability project*. Paper presented at the 4th New Directions in English Language Assessment Conference, Hanoi, Vietnam, October.

Field, J. (2013). Cognitive validity. In L. Taylor & A. Geranpayeh (Eds.), *Examining listening* (pp. 77–151). Cambridge: Cambridge University Press.

Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonisation. *Language Assessment Quarterly*, 1, 4, 253–266.

Geranpayeh, A. & Taylor, L. (Eds.) (2013). Examining listening: research and practice in assessing second language listening. *Studies in Language Testing 35*. Cambridge: Cambridge University Press.

Holzknecht, F., Eberharter, K., Kremmel, B., Zehentner, M., McCray, G., Konrad, E. & Spoettl, C. (2017). Looking into listening: using eye-tracking to establish the cognitive validity of the Aptis listening test. *ARAGs Research Reports Online, AR-G/2017/003.* London: British Council. (See: https://www.britishcouncil.org/exam/aptis/research/publications/non-technical-reports)

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

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, 21, 31–41.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50, 1–73.

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

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist,* 50(9), 741–749.

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

Nakatsuhara, F**.** (2014). *A Research Report on the Development of the Test of English for Academic Purposes (TEAP) Speaking Test for Japanese University Entrants – Study 1 & Study 2*. Retrieved from www.eiken.or.jp/teap/group/report.html

Nguyen, T. N. Quynh & Do, T. Minh (2015). *Developing a made-in-Vietnam standardised test of English proficiency for adults – The status quo and future development.* Paper presented at the International TESOL Symposium 2015, Da Nang, Vietnam.

Nguyen, T. N. Quynh, Nguyen, T.Q. Yen, Nguyen, T.P. Thao, Thai, H.L. Thuy, Bui, T. Sao & Carr, T. Nathan (2017). *Using multiple approaches to examine the dependability of the VSTEP speaking and writing assessments*. Paper presented at the 4th Asian Association for Language Assessment Conference, Taipei, Taiwan.

North, B., Martyniuk, W. & Panthier, J. (2010). Introduction: The manual for relating examinations to the Common European Framework of Reference for Languages in the context of the Council of Europe's work on language Education. In W. Martyniuk (Ed.) *Aligning Tests with the CEFR: Reflections on using the Council of Europe's draft Manual* (pp. 1–17). Cambridge: Cambridge University Press.

O'Sullivan, B. (2008). *City & Guilds Communicator IESOL Examination (B2) CEFR linking project: Case study.* Retrieved from: http://cdn.cityandguilds.com/ProductDocuments/International_English/General_English/8984/Additional_documents/8984_Case_study_v1.pdf

O'Sullivan, B. (2010). The City and Guilds Communicator examination linking project: a brief overview with reflections on the process. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual*. Cambridge: Cambridge University Press.

O'Sullivan, B. (2011). Language testing. In J. Simpson (Ed.), *Routledge handbook of applied linguistics*. Oxford: Routledge.

O'Sullivan, B. (2012). Assessment issues in languages for specific purposes. *Modern Language Journal,* 96, 71–88.

O'Sullivan, B. (2015a). Aptis test development approach. *Aptis Technical Report, TR/2015/001.* London: British Council. (See: https://www.britishcouncil.org/exam/aptis/research/publications/test-development-approach)

O'Sullivan, B. (2015b). Linking the Aptis reporting scales to the CEFR. *Aptis Technical Report, TR/2015/003.* London: British Council. (See: https://www.britishcouncil.org/exam/aptis/research/publications/reporting-scales)

O'Sullivan, B. (2017). *Establishing principles and procedures for linking examinations to the China's Standards of English*. Keynote address delivered at the 3rd International Conference on Language Testing and Assessment and the 5th New Directions in English Language Assessment Conference. Shanghai, China.

O'Sullivan, B. & Dunlea, J. (2015). Aptis General Technical Manual version 1.0. *Aptis Technical Report TR/2015/005.* London: British Council. (See: https://www.britishcouncil.org/aptis-general-technical-manual-version-10)

O'Sullivan, B. & Weir, C. (2011). Test development and validation. In B. O'Sullivan (Ed.), *Language testing: theories and practices* (pp. 13–32). Oxford: Palgrave Macmillan.

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

Taylor, L. (Ed.). (2012). *Examining speaking: Research and practice in assessing second language speaking*. Cambridge: Cambridge University Press.

Taylor, L. (2014). *A report on the review of test specifications for the reading and listening papers of the Test of English for Academic Purposes (TEAP) for Japanese University Entrants*. Retrieved from www.eiken.or.jp/teap/group/report.html

Taylor, L. & Galaczi, E. (2012). Scoring validity. In Taylor, L. (Ed.), *Examining speaking: research and practice in assessing second language speaking.* Cambridge: Cambridge University Press.

Tran, P., Nguyen, H., Dang, T., Nguyen, M., Nguyen, L., Huynh, T., Do, H. Nguyen, H. & Davidson, F. (2015). *A validation study on the newly developed Vietnam standardised English proficiency test.* Poster presented at the Language Testing Research Colloquium. Toronto, Canada.

Weir, C.J. (2005a). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing,* 22, 281–300.

Weir, C.J. (2005b). *Language Test Validation: an evidence-based approach*. Oxford: Palgrave.

Weir, C.J., Vidakovic, I. & Galaczi, E. (2013). *Measured constructs: a history of the constructs underlying Cambridge English language (ESOL) examinations 1913–2012*. Cambridge: Cambridge University Press.

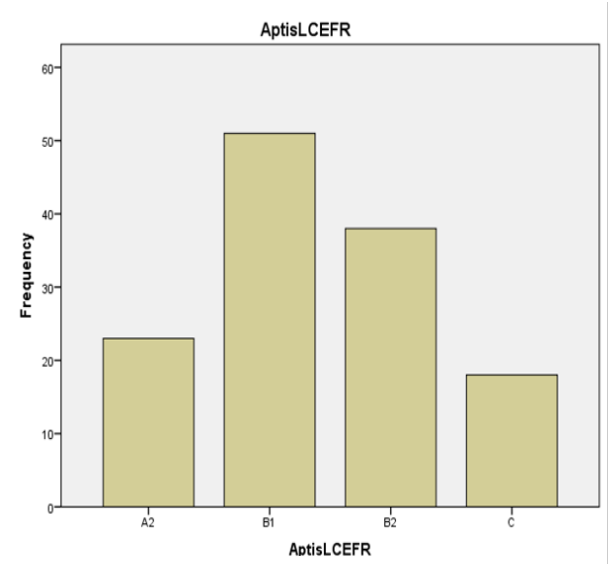
Wu, R. Y. F. (2014). Validating Second Language Reading Examinations: Establishing the Validity of the GEPT through Alignment with the Common European Framework of Reference, *Studies in Language Testing 41.* Cambridge: Cambridge University Press.

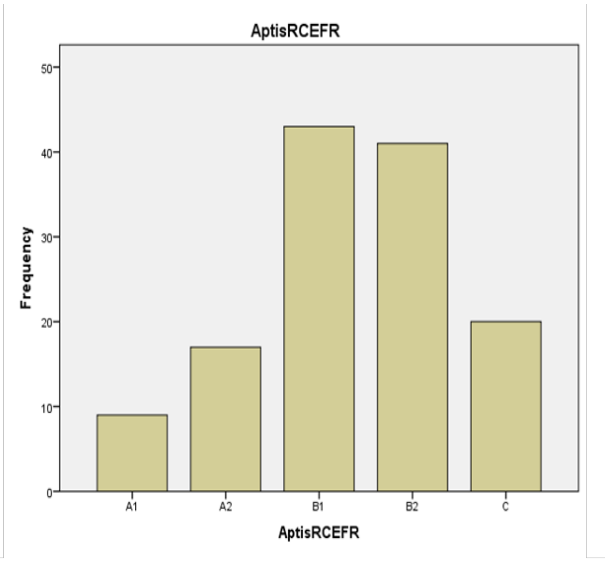
Wu, R., Yeh, H., Dunlea, J. & Spiby, R. (2016). Aptis-GEPT comparison study: Looking at two tests from multiple perspectives using the socio-cognitive model. *British Council Validations Series VS/2016/002.* London: British Council. (See https://www.britishcouncil.org/exam/aptis/research/publications/validation/aptis-gept-test-comparison-study)

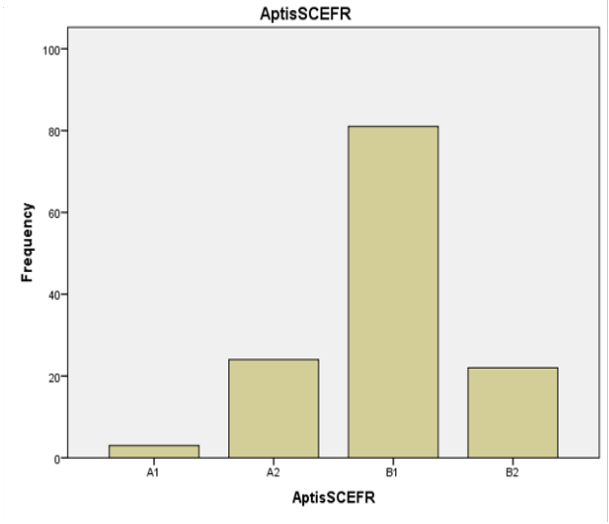
Appendix A:

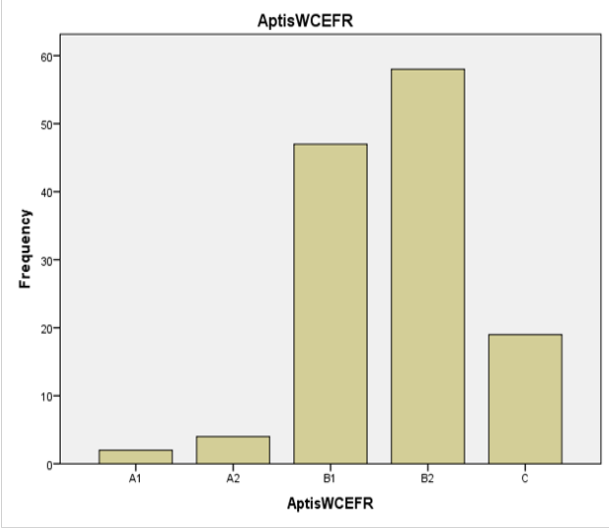
CEFR profiles for Aptis four skills components

CEFR level distributions for: Aptis Listening (top left); Aptis Reading (top right); Aptis Speaking (bottom left); and Aptis Writing (bottom right).







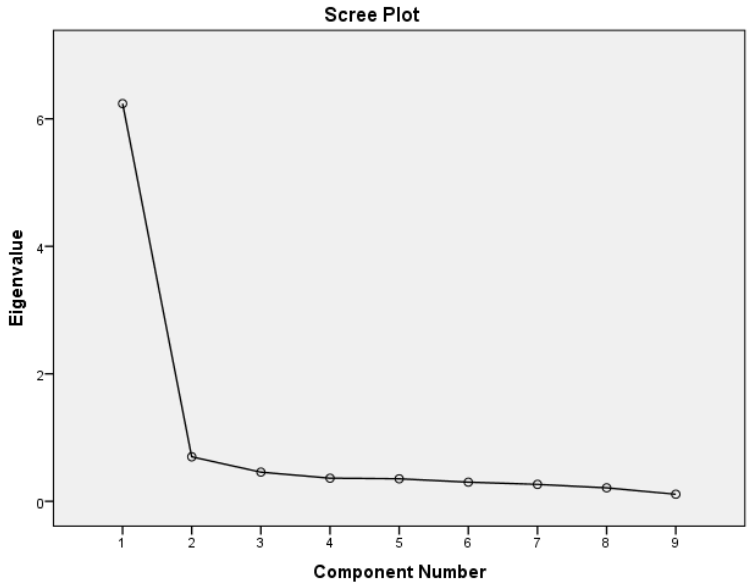


Appendix B:

Results of Principal Components Analysis

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.930
Bartlett's Test of Sphericity	Approx. Chi-Square	956.778
	df	36
	Sig.	.000



Component Matrix[a]

	Component 1
AptisGVScore	.921
AptisLScore	.829
AptisRScore	.816
AptisSScore	.881
AptisWScore	.853
VSTEPRScore	.803
VSTEPLScore	.669
VSTEPWScore	.870
VSTEPSScore	.827

Extraction Method: Principal Component Analysis.

Appendix C: Questionnaire data from pilot study

VSTEP

	Reading				Listening				Writing				Speaking			
	strongly agree	agree	disagree	strongly agree	agree	disagree	strongly disagree	strongly disagree	strongly agree	agree	disagree	strongly disagree	strongly agree	agree	disagree	strongly disagree
1. Generally speaking, the results of today's test are able to reflect my English ability.	10%	79%	8%	1%	7%	68%	21%	3%	19%	75%	4%	0%	12%	71%	8%	1%
2. The topics of the test are related to my life and/or working experiences.	12%	68%	17%	2%	8%	66%	21%	2%	20%	75%	2%	0%	15%	69%	7%	0%
3. I think the test instructions are clear.	25%	71%	2%	0%	22%	72%	4%	1%	25%	72%	1%	0%	20%	67%	6%	1%
4. I think the VSTEP test system provides a user-friendly interface.	24%	71%	4%	0%	20%	73%	6%	0%	25%	70%	3%	0%	19%	68%	5%	1%
5. I think the number of items of the VSTEP test is appropriate.	8%	60%	29%	2%	10%	67%	20%	1%	17%	75%	7%	0%	15%	69%	7%	1%
6. I think the time allotted for the VSTEP test is appropriate.	10%	61%	26%	2%	12%	64%	22%	1%	16%	69%	14%	1%	14%	66%	10%	1%
7. The orders of test parts and test items are appropriate.	16%	77%	4%	0%	16%	74%	5%	1%	20%	76%	2%	1%	15%	71%	6%	0%
8. Generally speaking, the sentence structures, vocabulary and phrases in the VSTEP test are commonly used in daily life/workplace.	11%	70%	16%	2%	11%	71%	15%	2%	18%	75%	6%	1%	16%	70%	7%	1%

	Reading				Listening				Writing				Speaking			
	strongly agree	agree	disagree	strongly agree	agree	disagree	strongly disagree	strongly disagree	strongly agree	agree	disagree	strongly disagree	strongly agree	agree	disagree	strongly disagree
9. Generally speaking, the speech rate of the listening input is appropriate.	—	—	—	—	7%	45%	40%	6%	—	—	—	—	—	—	—	—
10. The time intervals between test items in the listening test are appropriate.	—	—	—	—	9%	65%	21%	3%	—	—	—	—	—	—	—	—
11. Generally speaking, the accents of the listening input are clear and easy to understand.	—	—	—	—	7%	51%	33%	7%	—	—	—	—	—	—	—	—
12. The speech rate of the speaking examiners is appropriate.	—	—	—	—	—	—	—	—	—	—	—	—	17%	58%	5%	1%
13. The speaking examiner's pronunciation is clear and easy to understand.	—	—	—	—	—	—	—	—	—	—	—	—	17%	57%	5%	0%
14. The time for preparation in part 2 and 3 of the speaking test is appropriate.	—	—	—	—	—	—	—	—	—	—	—	—	14%	51%	14%	1%

	Very familiar	Quite familiar	Not very familiar	Totally not familiar
15. You are familiar with the VSTEP test format before taking the test.	3%	31%	48%	8%

	Very important	Quite important	Quite important	Not important
16. The influence of VSTEP result on your life/ work is…	48%	26%	14%	2%

Aptis Grammar and Vocabulary Test, Reading Test, and Listening Test

	Grammar and vocabulary				Reading				Listening			
	strongly agree	agree	disagree	strongly agree	agree	disagree	strongly disagree	strongly disagree	strongly agree	agree	disagree	strongly disagree
1. Generally speaking, the results of today's test are able to reflect my English ability.	29%	60%	6%	5%	25%	62%	6%	6%	30%	61%	5%	5%
2. The topics of the test are related to my life and/or working experiences.	36%	52%	7%	3%	33%	56%	7%	4%	30%	60%	6%	3%
3. I think the test instructions are clear.	49%	43%	1%	7%	45%	47%	1%	7%	46%	46%	2%	6%
4. I think the VSTEP test system provides a user-friendly interface.	48%	42%	3%	5%	43%	48%	4%	5%	43%	48%	2%	6%
5. I think the number of items of the VSTEP test is appropriate.	39%	49%	4%	7%	36%	54%	4%	6%	33%	55%	5%	5%
6. I think the time allotted for the VSTEP test is appropriate.	34%	52%	7%	7%	34%	55%	5%	6%	35%	55%	3%	7%
7. The orders of test parts and test items are appropriate.	40%	50%	3%	5%	32%	60%	2%	6%	36%	55%	1%	7%
8. Generally speaking, the sentence structures, vocabulary and phrases in the VSTEP test are commonly used in daily life/workplace.	32%	56%	5%	5%	31%	59%	4%	5%	37%	52%	5%	5%
9. Generally speaking, the speech rate of the listening input is appropriate.	—	—	—	—	—	—	—	—	33%	56%	5%	5%
10. Generally speaking, the accents of the listening input are clear and easy to understand.	—	—	—	—	—	—	—	—	31%	53%	12%	4%
11. The time intervals between test items in the listening test are appropriate.	—	—	—	—	—	—	—	—	35%	56%	3%	5%
12. I think listening to the test items twice enhances my performance.	—	—	—	—	—	—	—	—	46%	42%	4%	7%

Aptis Writing Test and Speaking Test

	Writing				Speaking			
	strongly agree	agree	disagree	strongly agree	agree	disagree	strongly disagree	strongly disagree
1. Generally speaking, the results of today's test are able to reflect my English ability.	24%	59%	12%	4%	25%	51%	18%	3%
2. The topics of the test are related to my life and/or working experiences.	30%	54%	10%	4%	28%	56%	11%	2%
3. I think the test instructions are clear.	44%	45%	5%	5%	35%	52%	7%	3%
4. I think the Aptis test system provides a user-friendly interface.	38%	50%	5%	6%	35%	48%	12%	1%
5. I think the number of items of the Aptis test is appropriate.	32%	48%	11%	7%	29%	50%	18%	1%
6. I think the time allotted for the Aptis test is appropriate.	26%	51%	14%	8%	26%	46%	22%	3%
7. The orders of test parts and test items are appropriate.	35%	53%	3%	6%	25%	61%	7%	2%
8. Generally speaking, the sentence structures, vocabulary and phrases in the Aptis test are commonly used in daily life/workplace.	34%	52%	7%	4%	31%	56%	9%	2%
9. Generally speaking, the speech rate of the recording of the Aptis Speaking Test is appropriate.	—	—	—	—	28%	53%	14%	3%
10. Generally speaking, the accents of the input in the Aptis Speaking Test are clear and easy to follow.	—	—	—	—	31%	56%	8%	3%
11. The time intervals between test items in the listening test are appropriate.	—	—	—	—	—	—	—	—
12. I think listening to the test items twice enhances my performance.	22%	39%	7%	5%	—	—	—	—
13. I prefer computer-based writing tests to paper-and-pencil-based writing tests.	33%	35%	18%	11%	—	—	—	—
14. I prefer face-to-face speaking tests to machine-recorded speaking tests.	—	—	—	—	33%	41%	14%	10%

Computer familiarity

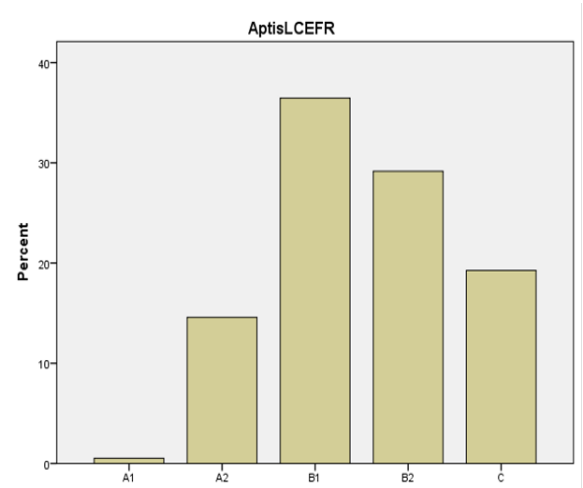
	YES		NO	
15. I have taken other computer-based English tests before taking today's Aptis test.	24%		75%	
16. I did the practice test online before taking today's Aptis test.	40%		59%	
	strongly agree	agree	disagree	strongly disagree
17. I often use computers.	56%	40%	1%	1%
18. I often read on computer.	50%	44%	4%	1%
19. I often type in English.	35%	47%	16%	1%
20. I usually listen to music or news on computers or digital devices.	46%	48%	3%	1%

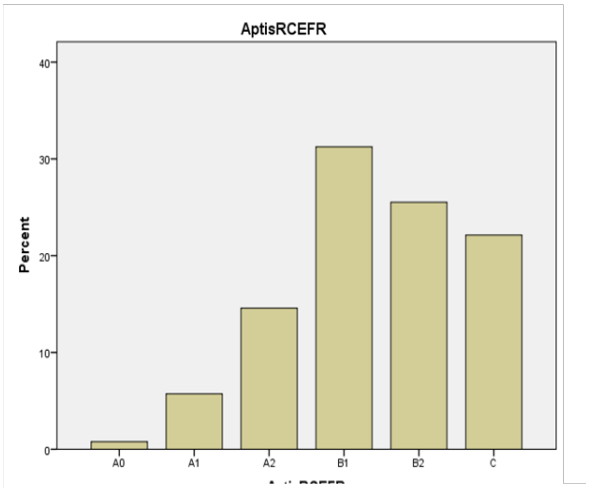
Aptis and VSTEP

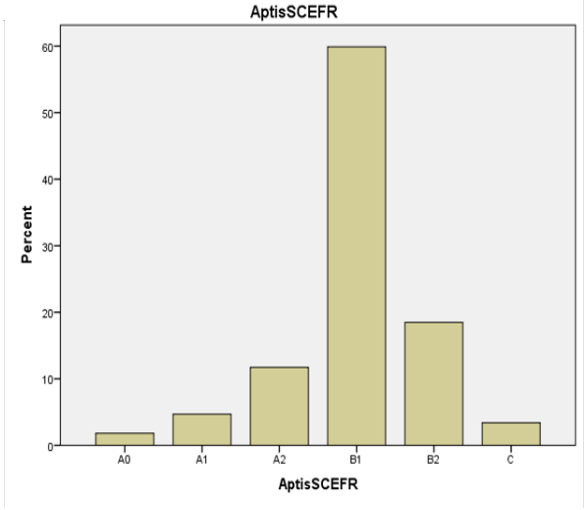
	Grammar and vocabulary		Reading		Listening		Writing		Speaking	
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
21. If you could, would you prefer to take the VSTEP on computer?			56%	41%	88%	8%	52%	44%	54%	40%
22. If you could, would you prefer to take the paper-and-pencil-based Aptis?	20%	73%	24%	69%	18%	76%	39%	56%		

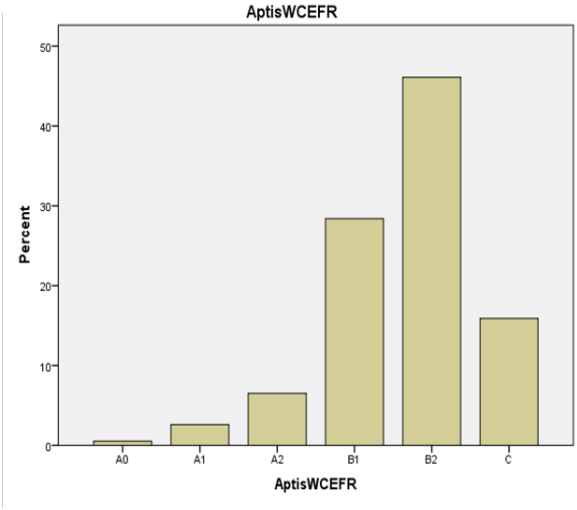
Appendix D: CEFR profiles for Aptis skills components in main study

CEFR level distributions for: Aptis Listening (top left); Aptis Reading (top right); Aptis Speaking (bottom left); and Aptis Writing (bottom right).





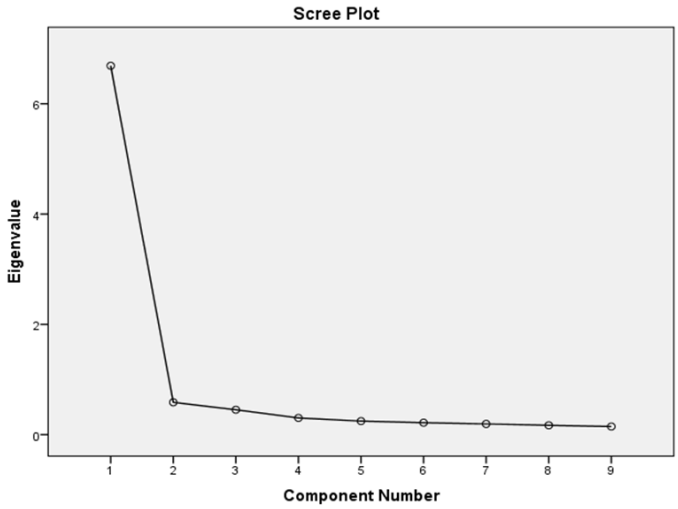




Appendix E:
Results of Principal Components Analysis

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.941
Bartlett's Test of Sphericity	Approx. Chi-Square	3374.184
	df	36
	Sig.	0.000



Component Matrix[a]	
	Component
	1
AptisGVScore	.921
AptisLScore	.841
AptisRScore	.850
AptisSScore	.857
AptisWScore	.842
VSTEPLScore	.858
VSTEPRScore	.891
VSTEPSScore	.860
VSTEPWScore	.836
Extraction Method: Principal Component Analysis.	
a. 1 components extracted.	

Appendix F: Questionnaire data for main study

VSTEP

	Reading				Listening				Writing				Speaking			
	strongly agree	agree	disagree	strongly agree	agree	disagree	strongly disagree	strongly disagree	strongly agree	agree	disagree	strongly disagree	strongly agree	agree	disagree	strongly disagree
1. Generally speaking, the results of today's test are able to reflect my English ability.	17%	74%	8%	0%	11%	71%	16%	1%	18%	71%	9%	1%	19%	71%	5%	0%
2. The topics of the test are related to my life and/or working experiences.	17%	73%	9%	0%	9%	68%	20%	1%	19%	73%	6%	0%	22%	67%	5%	0%
3. I think the test instructions are clear.	36%	61%	3%	0%	32%	64%	4%	0%	32%	62%	4%	0%	32%	59%	4%	0%
4. I think the VSTEP test system provides a user-friendly interface.	32%	63%	3%	1%	30%	64%	5%	0%	30%	63%	5%	1%	27%	63%	5%	0%
5. I think the number of items of the VSTEP test is appropriate.	16%	62%	20%	1%	16%	64%	17%	1%	17%	73%	8%	0%	21%	66%	8%	0%
6. I think the time allotted for the VSTEP test is appropriate.	16%	64%	17%	2%	14%	60%	21%	3%	13%	59%	24%	2%	20%	66%	9%	0%
7. The orders of test parts and test items are appropriate.	22%	71%	4%	1%	21%	68%	6%	1%	21%	70%	4%	1%	23%	67%	4%	0%
8. Generally speaking, the sentence structures, vocabulary and phrases in the VSTEP test are commonly used in daily life/workplace.	18%	69%	11%	1%	13%	69%	15%	1%	18%	73%	6%	0%	20%	70%	4%	0%

	Reading				Listening				Writing				Speaking			
	strongly agree	agree	disagree	strongly agree	agree	disagree	strongly disagree	strongly disagree	strongly agree	agree	disagree	strongly disagree	strongly agree	agree	disagree	strongly disagree
9. Generally speaking, the speech rate of the listening input is appropriate.	—	—	—	—	12%	55%	29%	3%	—	—	—	—	—	—	—	—
10. The time intervals between test items in the listening test are appropriate.	—	—	—	—	15%	58%	22%	3%	—	—	—	—	—	—	—	—
11. Generally speaking, the accents of the listening input are clear and easy to understand.	—	—	—	—	10%	53%	30%	4%	—	—	—	—	—	—	—	—
12. The speech rate of the speaking examiners is appropriate.	—	—	—	—	—	—	—	—	—	—	—	—	31%	57%	4%	0%
13. The speaking examiner's pronunciation is clear and easy to understand.	—	—	—	—	—	—	—	—	—	—	—	—	29%	57%	6%	1%
14. The time for preparation in part 2 and 3 of the speaking test is appropriate.	—	—	—	—	—	—	—	—	—	—	—	—	20%	58%	12%	1%

	Very familiar	Quite familiar	Not very familiar	Totally not familiar
15. You are familiar with the VSTEP test format before taking the test.	5%	23%	37%	33%

	Very important	Quite important	Quite important	Not important
16. The influence of VSTEP result on your life/ work is…	26%	34%	31%	9%

Aptis Grammar and Vocabulary Test, Reading Test, and Listening Test

	Grammar and vocabulary				Reading				Listening			
	strongly agree	agree	disagree	strongly agree	agree	disagree	strongly disagree	strongly disagree	strongly agree	agree	disagree	strongly disagree
1. Generally speaking, the results of today's test are able to reflect my English ability.	26%	59%	11%	4%	23%	60%	13%	4%	28%	56%	11%	5%
2. The topics of the test are related to my life and/or working experiences.	23%	61%	12%	4%	23%	61%	12%	3%	28%	58%	9%	4%
3. I think the test instructions are clear.	51%	39%	4%	6%	49%	38%	6%	5%	51%	39%	3%	6%
4. I think the VSTEP test system provides a user-friendly interface.	54%	38%	2%	6%	49%	41%	4%	6%	51%	40%	3%	5%
5. I think the number of items of the VSTEP test is appropriate.	37%	50%	6%	6%	36%	50%	8%	5%	39%	45%	8%	6%
6. I think the time allotted for the VSTEP test is appropriate.	36%	51%	7%	6%	36%	50%	7%	7%	35%	50%	8%	6%
7. The orders of test parts and test items are appropriate.	36%	52%	5%	6%	38%	51%	4%	6%	38%	50%	4%	6%
8. Generally speaking, the sentence structures, vocabulary and phrases in the VSTEP test are commonly used in daily life/workplace.	28%	55%	11%	5%	26%	57%	10%	5%	29%	59%	6%	5%
9. Generally speaking, the speech rate of the listening input is appropriate.	—	—	—	—	—	—	—	—	36%	50%	8%	4%
10. Generally speaking, the accents of the listening input are clear and easy to understand.	—	—	—	—	—	—	—	—	35%	45%	14%	4%
11. The time intervals between test items in the listening test are appropriate.	—	—	—	—	—	—	—	—	35%	52%	6%	6%
12. I think listening to the test items twice enhances my performance.	—	—	—	—	—	—	—	—	47%	39%	5%	8%

Aptis Writing Test and Speaking Test

	Writing				Speaking			
	strongly agree	agree	disagree	strongly agree	agree	disagree	strongly disagree	strongly disagree
1. Generally speaking, the results of today's test are able to reflect my English ability.	25%	56%	14%	5%	19%	52%	14%	4%
2. The topics of the test are related to my life and/or working experiences.	24%	58%	11%	5%	21%	52%	11%	4%
3. I think the test instructions are clear.	47%	41%	4%	6%	35%	44%	4%	5%
4. I think the Aptis test system provides a user-friendly interface.	46%	40%	8%	4%	32%	48%	4%	4%
5. I think the number of items of the Aptis test is appropriate.	28%	52%	13%	5%	25%	51%	6%	5%
6. I think the time allotted for the Aptis test is appropriate.	27%	49%	18%	6%	21%	42%	19%	7%
7. The orders of test parts and test items are appropriate.	32%	55%	6%	5%	24%	52%	6%	4%
8. Generally speaking, the sentence structures, vocabulary and phrases in the Aptis test are commonly used in daily life/workplace.	25%	61%	8%	5%	23%	54%	6%	5%
9. Generally speaking, the speech rate of the recording of the Aptis Speaking Test is appropriate.	—	—	—	—	24%	50%	7%	5%
10. Generally speaking, the accents of the input in the Aptis Speaking Test are clear and easy to follow.	—	—	—	—	24%	51%	7%	5%
11. The time intervals between test items in the listening test are appropriate.	—	—	—	—	—	—	—	—
12. I think listening to the test items twice enhances my performance.	—	—	—	—	—	—	—	—
13. I prefer computer-based writing tests to paper-and-pencil-based writing tests.	31%	32%	23%	11%	—	—	—	—
14. I prefer face-to-face speaking tests to machine-recorded speaking tests.	—	—	—	—	31%	37%	16%	10%

Computer familiarity

	Yes		No	
15. I have taken other computer-based English tests before taking today's Aptis test.	29%		66%	
16. I did the practice test online before taking today's Aptis test.	23%		72%	
	strongly agree	agree	disagree	strongly disagree
17. I often use computers.	53%	38%	4%	1%
18. I often read on computer.	50%	36%	8%	0%
19. I often type in English.	44%	37%	13%	1%
20. I usually listen to music or news on computers or digital devices.	55%	36%	4%	1%

Aptis and VSTEP

	Grammar and vocabulary		Reading		Listening		Writing		Speaking	
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
21. If you could, would you prefer to take the VSTEP on computer?	—	—	64%	30%	86%	8%	58%	35%	47%	44%
22. If you could, would you prefer to take the paper-and-pencil-based Aptis?	34%	57%	42%	51%	28%	63%	41%	51%	—	—

Appendix G: Example of construct definition template for long reading task

Categories Reading	Task 1	(Task 1) Item 1	(Task 1) Item 2	(Task 1) Item 3	(Task 1) Item 4	(Task 1) Item 5
	CONSENSUS	CONSENSUS	CONSENSUS	CONSENSUS	CONSENSUS	CONSENSUS
Features of the TASK	**Features of the TASK**	**Features of the TASK**	**Features of the TASK**	**Features of the TASK**	**Features of the TASK**	**Features of the TASK**
Skill focus	sentence comprehension, lexis					
Task Level (CEFR)	A1					
Response format	Multiple choice gap fill					
Items per task	5					
Cognitive processing 1	Careful reading: local					
Cognitive processing 2	Establishing propositional meaning (cl./sent. level)					
Content knowledge	1 (General)					
Cultural specificity	1 (Neutral)					
Features of the Input Text	**Features of the Input Text**	**Features of the Input Text**	**Features of the Input Text**	**Features of the Input Text**	**Features of the Input Text**	**Features of the Input Text**
Domain	Personal					
Discourse mode	Descriptive					
Nature of information	Only concrete					
Topic	Daily life					
Text genre	Personal letters / e-mail					
Presentation	Verbal (written)					
Features of the Response	**Features of the Response**	**Features of the Response**	**Features of the Response**	**Features of the Response**	**Features of the Response**	**Features of the Response**
Key information		Within Sentences	Within Sentences	Within Sentences	Within Sentences	Within Sentences
Operation		Main idea / conclusions	Main idea / conclusions	Main idea / conclusions	Main idea / conclusions	Main idea / conclusions
Question presentation		Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)
Option Presentation		Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)	Verbal (written)

Appendix H: Construct definition for final long reading task in Aptis by LTRGI

Categories Reading	APTIS Task 4
	CONSENSUS
Features of the TASK	**Features of the TASK**
Skill focus	paragraph comprehension, reading for gist, understanding main ideas of longer complex text
Task Level (CEFR)	B2
Response format	Matching headings to text
Items per task	7
Cognitive processing 1	Expeditious reading: global
Cognitive processing 2	Building a mental model
Content knowledge	2
Cultural specificity	1 (Neutral)

Features of the Response	(Task 4) Item 1	(Task 4) Item 2	(Task 4) Item 3	(Task 4) Item 4	(Task 4) Item 5	(Task 4) Item 6	(Task 4) Item 7
Key information	across sentences	across sentences	across sentences	across sentences	across sentences	across sentences	across sentences
Operation	Main idea / conclusions	Main idea / conclusions	Main idea / conclusions	Main idea / conclusions	Main idea / conclusions	Main idea / conclusions	Main idea / conclusions

Appendix I: Construct definition for final long reading task in Aptis by ULIS team

Categories Reading	APTIS Task 4
	CONSENSUS
Features of the TASK	**Features of the TASK**
Skill focus	skimming for main ideas of paragraph
Task Level (CEFR)	B2
Response format	Matching headings to text
Items per task	7
Cognitive processing 1	**Expeditious reading: global**
Cognitive processing 2	**Building a mental model**
Content knowledge	4
Cultural specificity	1 (Neutral)

Features of the Response	(Task 4) Item 1	(Task 4) Item 2	(Task 4) Item 3	(Task 4) Item 4	(Task 4) Item 5	(Task 4) Item 6	(Task 4) Item 7
Key information	across sentences	across sentences	across sentences	across sentences	across sentences	across sentences	across sentences
Operation	Gist	Gist	Gist	Gist	Gist	Gist	Gist

Appendix J: Construct definition for final long reading task in Aptis by ULIS team

Categories Reading	Task 4
	CONSENSUS
Features of the TASK	**Features of the TASK**
Skill focus	identifying main ideas, finer details and implied relationships, understanding longer complex texts
Task Level (CEFR)	C1
Response format	MCQ
Items per task	10
Cognitive processing 1	Careful reading: global
Cognitive processing 2	Creating a text level representation (disc. structure)
Content knowledge	2
Cultural specificity	2

Categories Reading	(Task 4) Item 1	(Task 4) Item 2	(Task 4) Item 3	(Task 4) Item 4	(Task 4) Item 5	(Task 4) Item 6	(Task 4) Item 7	(Task 4) Item 8	(Task 4) Item 9	(Task 4) Item 10
Key information	Within sentences	across sentences	Within sentences	across sentences	across sentences	Within sentences	across sentences	across sentences	across paragraphs	across sentences
Operation	Specific information	Main idea / conclusions	Specific information	Main idea / conclusions	Main idea / conclusions	Main idea / conclusions	Opinion	Main idea / conclusions	Test structure / connections between the parts	Opinion

Appendix K: Construct definition for long reading task in VSTEP by ULIS team

Categories Reading	Task 4
	CONSENSUS
Features of the TASK	**Features of the TASK**
Skill focus	Reading for gist, main ideas and specific information
Task Level (CEFR)	C1
Response format	MCQ
Items per task	10
Cognitive processing 1	**Careful reading: global**
Cognitive processing 2	**Creating a text level representation (disc. structure)**
Content knowledge	5 (Specific
Cultural specificity	1 (Neutral)

Categories Reading	(Task 4) Item 1	(Task 4) Item 2	(Task 4) Item 3	(Task 4) Item 4	(Task 4) Item 5	(Task 4) Item 6	(Task 4) Item 7	(Task 4) Item 8	(Task 4) Item 9	(Task 4) Item 10
Key information	Within sentences	across paragraphs	Within sentences	Within sentences	across paragraphs	Within sentences	across sentences	Within sentences	across paragraphs	across paragraphs
Operation	Specific information	Main idea / conclusions	Specific information	Main idea / conclusions	Main idea / conclusions	Main idea / conclusions	Opinion	Main idea / conclusions	Test structure / connections between the parts	Opinion

British Council Assessment Research Group

The Assessment Research Group was formed in 2013 to support the British Council's work in assessment and testing across the world. The team is responsible for ensuring that all new assessment products and new uses of existing products are supported by the most up-to-date research. They also continuously evaluate the quality of British Council assessment products.

APTIS-VSTEP COMPARABILITY STUDY: INVESTIGATING THE USAGE OF TWO EFL TESTS IN THE CONTEXT OF HIGHER EDUCATION IN VIETNAM

VS/2018/001

J. Dunlea, R. Spiby, T.N. Quynh Nguyen, T.Q. Yen Nguyen, T.M. Huu Nguyen, T.P. Thao Nguyen, H.L. Thuy Thai and T. Sao Bui

www.britishcouncil.org/aptis/research

BRITISH COUNCIL VALIDATION SERIES

Published by British Council
10 Spring Gardens
London SW1A 2BN