CAMBRIDGE ENGLISH
Language Assessment
Part of the University of Cambridge

100 CAMBRIDGE ENGLISH
CENTENARY 1913–2013

Research Notes

Issue 55

February 2014

ISSN 1756-509X

CAMBRIDGE ENGLISH
Language Assessment
Part of the University of Cambridge

Research Notes

Issue 55 / February 2014
A quarterly publication reporting on learning, teaching and assessment

Printed in the United Kingdom by Canon Business Services

Research Notes

Issue 55

February 2014

Contents

Editorial

Welcome to issue 55 of *Research Notes*, our quarterly publication reporting on research matters relating to learning, teaching and assessment within Cambridge English Language Assessment. This issue draws the attention of the *Research Notes* readership to the exams and research being conducted at Cambridge Michigan Language Assessments (CaMLA).

In September 2010, Cambridge English Language Assessment and the University of Michigan English Language Institute Testing and Certification Division formed CaMLA. This new joint venture brought together two organisations that have long histories in language education creating opportunities for collaboration. An important benefit of this partnership is our ability to offer stakeholders a wider range of exams to meet their specific needs.

The articles in this issue focus on fairness principles, frameworks designed to facilitate item writing, and exam revision and validation studies. Each article gives the reader a better understanding of CaMLA's suite of tests as well as the processes and concerns that are specific to their context of use.

The issue begins with an article describing the principles and procedures CaMLA and Cambridge English Language Assessment follow to ensure their tests are fair for all candidates. Katie Weyant and Amanda Chisholm show how fairness is embedded in all stages of the test development processes of both organisations even though the specific procedures and the way they are operationalised differ slightly in response to their unique operational contexts. In many ways, both organisations have embraced Spaan's (2000:35) proposal that test developers enter into a 'social contract' with stakeholders, which entails ensuring their tests are fair, accurate and valid as well as providing users with transparent and comprehensive information about all aspects of their tests (see Cambridge English Language Assessment 2013). This article highlights the link between test fairness and test validity in that concerns over fairness will inevitably lead to concerns about the validity and reliability of test results.

The next article by Fabiana MacMillan, Daniel Walter and Jessica O'Boyle deals with the perennial question of what makes an item more or less difficult. As the authors point out, previous research has shown that item writers are not particularly accurate when it comes to predicting item difficulty (see Fortus, Coriat and Fund 1998, Hamp-Lyons and Mathias 1994 and Sydorenko 2011); however, being able to identify factors that affect difficulty can facilitate item writing and allow test developers to better manage their item banks. In response to this issue, MacMillan, Walter and O'Boyle describe the development of a framework for predicting the difficulty of grammatical items being produced for the Examination for the Certificate of Proficiency in English (ECPE™) test, which is targeted at Level C2 of the Common European Framework of Reference (CEFR) (Council of Europe 2001). The authors make use of corpora and second language acquisition research to create a tagging rubric that can be applied to items testing grammatical knowledge and use. This rubric was then trialled and the predictive ability of it was tested. This study highlights the challenges involved in trying to systematically classify C2 level grammar items but also points to the need for more research into the criterial features that distinguish one level from another, in particular at the C levels.

The Examination for the Certificate of Competency in English (ECCE™), which is aimed at the B2 level of the CEFR, was recently revised and the next two articles focus on the changes made to the reading and listening papers respectively. First, Fabiana MacMillan, Mark Chapman and Jill Rachele Stucker describe a new task type introduced into the reading paper and the rationale for its inclusion. This task involves candidates having to process information from more than one text in order to answer a series of comprehension questions. It requires readers to develop an intertextual representation, which is a trait associated with high-level readers (Khalifa and Weir 2009). MacMillan, Chapman and Stucker present the theoretical basis for including this task type, the challenges associated with developing these reading tasks as well as candidate performance. Although they found that writing these tasks proved to be quite challenging, they perform as expected and improve the construct coverage of the paper. The revision process involved in updating the listening paper is then described by Natalie Nordby Chen and Jayanti Banerjee. In response to stakeholder consultative exercises, a review of the B2 listening CEFR descriptors and current research into second language listening, Part 2 of the listening paper was modified from a single long interview to four short talks, which allowed the paper to cover more listening constructs. Both of these articles provide a good example of a rational and explicit approach to test revision and development.

The last three articles focus on the validation of different CaMLA tests. The first article in this section investigates the equivalence of writing prompts in the ECPE test, a C2 level exam. As this writing test offers candidates a choice between two writing prompts, Mark Chapman, Crystal Collins, Barbara Allore Dame and Heather Elliott investigate which prompt features should be controlled for to ensure that the writing output is also comparable. This is important as it links to the issue of fairness: if the features and characteristics of the writing produced are shaped by the prompt, then comparable prompts will ensure that scripts can be scored consistently. They found that when domain, length and task wording were kept consistent, candidate output did not vary substantially. Their recommendations could lead to a more systematic approach to prompt development and ensure not only comparability within a test but across test versions.

Ildiko Porter-Szucs and Ummehaany Jameel then present research that focuses on the International Teaching Assistant Speaking Assessment (ITASA™) which is used to evaluate the language proficiency of international teaching assistants (ITAs). As the test is designed to replicate, as much as possible, the situations and settings that ITAs operate within, the authors were interested in finding out the extent to which the 'Office-Hour Role Play' (Task 3) elicits nativelike formulaic sequences and whether the presence of these

features in candidate responses influences their scores. Although they didn't find any relationship between the use of formulaic language and candidate ratings, they were able to demonstrate that the language produced by candidates was similar to that produced by native speakers performing similar tasks. This study provides further support that this test is fit for purpose.

The final article describes an investigation into the underlying structure of the English Placement Test (EPT™) by using Confirmatory Factor Analysis (CFA). The EPT was designed to measure overall receptive language ability rather than the individual skills and language elements underlying this ability, which is why scores are reported using a single scale. In light of this, Daniel Walter and Jasmine Hentschel investigate whether the different components of the EPT are assessing unique aspects of language proficiency or overall receptive ability in order to justify their reporting practices. They found that the best fitting model had grammar linked to both reading and listening, which would make it difficult to separate the components into sub-scores, thus supporting their position that EPT measures general receptive language ability.

The articles in this issue demonstrate the range of activities the CaMLA team are engaged in to ensure their assessment products are fit for purpose. We also hope this issue shows how two organisations which operate in different contexts, and which have longstanding operational traditions, can both work together and learn from each other in order to continuously improve their assessments and ensure that essential test qualities are met and delivered according to stakeholder expectations.

Cambridge English Language Assessment (2013) *Principles of Good Practice: Quality Management and Validation in Language Assessment*, available online: www.cambridgeenglish.org/research-and-validation/quality-and-accountability

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching and Assessment*, Cambridge: Cambridge University Press.

Fortus, R, Coriat, R and Fund S (1998) Prediction of item difficulty in the English subtest of Israel's inter-university psychometric entrance test, in Kunnan, A (Ed) *Validation in Language Assessment: Selected Papers From the 17th Language Testing Research Colloquium, Long Beach,* Mahwah: Lawrence Erlbaum, 61–87.

Hamp-Lyons, L and Mathias, S P (1994) Examining expert judgments of task difficulty on essay tests, *Journal of Second Language Writing* 3 (1), 49–68.

Khalifa, H and Weir, C J (2009) *Examining Reading: Research and Practice in Assessing Second Language Reading*, Studies in Language Testing volume 29, Cambridge: UCLES/Cambridge University Press.

Spaan, M (2000) Enhancing fairness through a social contract, in Kunnan, A (Ed) *Fairness and Validation in Language Assessment*, Studies in Language Testing volume 9, Cambridge: UCLES/Cambridge University Press, 35–38.

Sydorenko, T (2011) Item writer judgments of item difficulty versus actual item difficulty: A case study, *Language Assessment Quarterly* 8 (1), 34–52.

Safeguarding fairness principles through the test development process: A tale of two organisations

KATIE WEYANT ASSESSMENT GROUP, CAMBRIDGE MICHIGAN LANGUAGE ASSESSMENTS, USA
AMANDA CHISHOLM ASSESSMENT AND OPERATIONS GROUP, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

Introduction

In the past several decades, professionals and scholars in the field of language assessment have become increasingly aware of the need to define a code of ethics; one that not only guides professional conduct, but also applies to aspects of tests such as fairness and equality (McNamara and Roever 2006). As McNamara and Roever (2006) point out, in addition to using psychometric tools to ensure the fairness of a test, testing organisations have also put policies, guidelines, and fairness reviews in place in order to safeguard fairness principles throughout the test development process.

It must be acknowledged from the outset that fairness is a slippery concept, and can be discussed in numerous testing-related contexts, e.g. equal opportunities, test security, special arrangements for test takers with disabilities (Cambridge English Language Assessment 2013a:8–9), or in terms of principles, e.g. Fairness as a Lack of Bias, Fairness as Equitable Treatment in the Testing Process, Fairness as Equality in Outcomes of Testing, and Fairness as Opportunity to Learn, outlined by the *Standards for Educational and Psychological Testing* prepared by the American Educational Research Association, American Psychological Association and the National Council on Measurement in Education (1999:73–77). As Hamp-Lyons acknowledges, 'fairness' is a difficult concept because from no one standpoint can a test be viewed as 'fair' or 'not fair' (Hamp-Lyons 2000:32), placing fairness in the context of fairness to candidates, teachers, and other stakeholder groups.

This article focuses specifically on fairness as the avoidance of construct-irrelevant variance in the test development process, and tools used by Cambridge English Language Assessment and Cambridge Michigan Language Assessments (CaMLA) to implement safeguards against bias, both in terms of content-related and response-related sources of bias. While both organisations share the goal of creating

fair and unbiased assessments, the differing approaches taken by each organisation have been shaped by their unique histories and practical concerns. The principle of safeguarding fairness is embedded throughout the processes and policies of both organisations, but for the purposes of this article, we will tease out the numerous ways in which fairness is ensured.

Principles

The CaMLA Fairness Committee is responsible for ensuring that the six CaMLA Fairness Principles outlined below are adhered to during item development. The committee reviews, discusses, and updates CaMLA's *Fairness Overview* and *Fairness Principles* documents, answers enquiries from item developers on specific fairness concerns, and conducts training for staff and item writers. Face-to-face training sessions are conducted for internal staff, and external freelance item writers are provided with self-guided training modules that are delivered via a secure online file-sharing platform. All training sessions include an overview of the Fairness Principles and procedures, as well as practice exercises (e.g. a task in which one reads an item and determines whether or not it violates one or more of the Fairness Principles, which principle(s) it violates, and why).

The six principles that comprise CaMLA's Fairness Principles include:

- **Demonstrate Respect for People**, e.g. by attending to people-group origins with sensitivity; avoiding implying the superiority of gender, race, nation, or social group; avoiding perpetuating or accepting stereotypes; using gender-neutral labels and pronouns, if possible; balancing across a test a diverse representation of individuals in authoritative positions.
- **Demonstrate Respect for Personal Convictions and Beliefs**, including religious beliefs and holidays.
- **Demonstrate Sensitivity to Population Differences and World Knowledge**, e.g. by not assuming that there is only one perspective on or interpretation of facts; not assuming familiarity with specific technology; not assuming familiarity with North American customs; not assuming that 'America' refers only to the United States; avoiding elitist language and topics.
- **Avoid Undue Negativity**, such as academic failure, achievement gaps, and difficult testing situations; unnecessary meanness; addictive or harmful behaviours; crime and incarceration.
- **Avoid Unduly Controversial or Upsetting Topics**, such as prompting test takers to criticise or praise national practices, specific politicians or controversial political debates; war, violence, suicide; poverty, income disparity; famine, disease, death, dying, terminal illness; graphic or upsetting medical procedures; preference to avoid intimate situations.
- **Avoid Testing Construct-irrelevant Knowledge**, such as maths or science.

Corresponding principles are detailed at Cambridge English Language Assessment within the item writer guidelines (IWGs), rather than a specific 'fairness principles' document. When item writers submit potential examination material, it is reviewed against the IWGs. Material may be rejected or feedback given because it does not satisfy the requirements of the IWGs. The approach is largely topic-based, detailing suitable and unsuitable topics, and providing other topic-related considerations. This approach came out of a concern to avoid test anxiety, i.e. disadvantaging candidates by potentially distressing them, and is one way of standardising input from item writers. In the *Item Writer Guidelines: Information Common to All Papers, FCE, CAE and CPE* (now known as *Cambridge English: First; Cambridge English: Advanced* and *Cambridge English: Proficiency*) (Cambridge ESOL 2006:7) the following topics are judged unsuitable for all Cambridge English Language Assessment tests:

- alcohol
- cigarettes (where smoking is the focus of a text, picture or task)
- drugs
- examinations, passing and failing
- gambling
- historical subjects or references likely to offend certain nations or groups
- national standpoints (in particular those where the practices of one country may be perceived negatively by others)
- politics
- potentially distressing topics (e.g. death, terminal illness, severe family/social problems, natural disasters and the objects of common phobias such as spiders or snakes)
- religion (including related topics such as those which are not acceptable to certain religions)
- sex, sexuality and bodily functions
- stereotypes (includes racism, sexism, cultural clichés and attitudes which could be interpreted as patronising towards other countries, cultures, beliefs or individuals)
- war.

These IWGs, echoing CaMLA's principles of avoiding undue negativity and unduly controversial or upsetting topics, specify that examination material 'must not contain anything that might upset or distract candidates as this will affect their performance', with the caveat that 'common sense is essential in interpreting these unsuitable topics' (Cambridge ESOL 2006:7). The IWGs state that it might be acceptable for some of the topics above to be mentioned in a text, but that they may not be suitable as the main focus of material. For instance, these IWGs specify that a text focusing on British drinking habits may be offensive to some candidates, but that a mention of a glass of wine may be acceptable in texts, but not in visuals. The IWGs clarify that it is not the topic per se that is unsuitable, but the treatment of it in any given text, and that 'judgement may need to be the deciding factor' (Cambridge ESOL 2006:8). Such judgements relate to the candidature, level and purpose of each test. For instance, English for Specific Purposes (ESP) tests by necessity have particular approaches to fairness issues. The *Item Writer Guidelines: Information*

Common to All Papers, ILEC (now known as *Cambridge English: Legal*) warns that 'sensitive legal topics such as employment law relating to redundancy and harassment in the workplace may be suitable, but should be approached with caution' (Cambridge ESOL 2007:7). It is also important to note that judgements take place at a certain point in time, and that the list of unsuitable topics is regularly reviewed, to ensure that it is keeping pace with changing social norms.

The *Item Writer Guidelines: Information Common to All Papers, FCE, CAE and CPE* detail other considerations in topic selection, such as general knowledge, specialised material and cultural context. Similar in nature to CaMLA's principle of avoiding testing construct-irrelevant knowledge, IWGs call upon writers to avoid items that test general knowledge, and to avoid material that would be too specialised for most candidates to understand. The IWG concept of cultural context is similar in meaning to the CaMLA principle of demonstrating sensitivity to population differences and world knowledge.

Other main forces that shape the current practices of CaMLA and Cambridge English Language Assessment are practical considerations familiar to most language testing organisations, such as those involved in the logistics of item writer training.

External item writers

In both organisations, item writers are given training on understanding and using fairness principles. At Cambridge English Language Assessment, such principles are embedded in long-term item writer training. They underpin the initial application process, in which potential item writers are given the topic-related information above, and are assessed on their ability to pick out aspects of authentic texts which would be inappropriate for examination material.

Throughout their initial training commissions, writers are evaluated on their sensitivity to cultural issues. After that, writers will join test-specific teams, and receive face-to-face training on how to write for those tests. Teams are headed by Chairs, external consultants who are responsible for the content of each test, who give new writers additional support in their first year, e.g. reviewing potential source texts for cultural context issues before new writers begin crafting items. Writers also learn from their peers in face-to-face editing meetings led by the Chair, where a group of writers review their own material. When this material has been pretested or trialled with candidates, writers are given individual feedback on the performance of their items as well as feedback from the pretesting centre related to the appropriacy of the topics.

The Cambridge English Language Assessment item writer training model has the luxury of various face-to-face training sessions, as most of the item writers live in the UK, where distances are relatively short, but this is not as practical in the USA. Thus the CaMLA approach utilises the convenience of modern technology. CaMLA has developed two Fairness Training Modules, available on a secure online platform to all external and internal staff members. The first of the two includes an introduction to the Fairness Principles, an opportunity to identify items in violation of the Fairness Principles and how to revise them when possible, and a test that certifies them to perform fairness reviews on multiple-choice items. The second CaMLA Fairness Training Module demonstrates how the Fairness Principles can be adapted and applied to constructed response assessments (i.e. writing and speaking tests). It also has three stages that allow the trainee to apply the principles, perform practice reviews, and become certified to perform fairness reviews on constructed response item types.

Cambridge English Language Assessment is also moving towards internet-based self-access training, with four core modules currently under development, targeting each of the four skills. More specialist modules are planned for the future. This would be an area where CaMLA and Cambridge English Language Assessment could pool expertise, and produce joint self-access training modules. Both organisations take an important step towards reducing unintended construct-irrelevant variance by training external item writers to consistently recognise and avoid potential fairness and bias issues.

Item developers and development

All internal CaMLA Assessment and Multimedia staff members, in addition to external item writers and reviewers, are fully trained to determine if items adhere to CaMLA Fairness Principles and benefit from face-to-face discussions with members of the Fairness Committee. Although there is a designated Fairness Review stage in the item development path in which a reviewer scans for violations of the Fairness Principles, each staff member who comes into contact with items is trained to look for fairness violations. Each reviewer brings a fresh perspective to items; what seems harmless to one reviewer might be clearly offensive to another.

Cambridge English Language Assessment, similarly, has fairness-related checks throughout the test production process. New staff members are introduced to the test production cycle, which includes a shadowing programme, so that fairness issues can be discussed in the context of actual examination material in face-to-face meetings. Fairness issues can be a reason for rejecting tasks at the pre-editing stage, e.g. a text about husbands and wives who worked together was rejected by the *Cambridge English: Advanced* listening team, on the grounds that the interest of the piece lay in whether the husbands and wives argued or not. At the editing stage, fairness issues also inform discussion of items. For instance, on speaking tests, when using photos, any alcohol in a photo would be turned into juice.

CaMLA and Cambridge English Language Assessment have a similar approach to fairness issues when recording audio for listening tests. Ease of distinguishing voices is considered. As the IWGs for Cambridge listening papers state, it is important that candidates can easily distinguish between voices, so dialogues are recorded with one male and one female voice. The IWGs specify that in tasks involving three speakers, the voices should be disambiguated through the use of names, or by making sure that a female interviewer speaks to a male interviewee first.

Fairness issues are also included when creating artwork for

tests. At CaMLA, staff on the in-house graphic design team are trained fairness reviewers. The CaMLA Fairness Principles are adhered to when creating artwork, and images are altered if deemed unfair at a later review stage. At Cambridge English Language Assessment, fairness issues underpin artwork specifications, outlined in general guidelines for artists working on tests with a broad international candidature (Cambridge ESOL 2010:4):

- do not include pets
- do not include alcohol
- females should be depicted wearing capped sleeves or long sleeves and three-quarter-length skirts or long trousers
- people should not be touching each other
- the occasional character in groups should be Asian/black, which would reflect most candidates' experience in their home countries
- images should not be obviously British
- swimming pools should show swimmers with heads and one arm doing the crawl only
- commonly-used safety equipment should be shown where appropriate, e.g. people riding in cars should be wearing seatbelts.

Trialling constructed response items

Though trained writers and reviewers scan each item for potential fairness and bias issues throughout the development process, even the most seemingly innocuous item could have unintended negative consequences when administered live. This is particularly true with constructed response items, i.e. speaking and writing prompts. Therefore, both organisations have incorporated important steps in their current practices to reduce the risk of inadvertently introducing construct-irrelevant variance in these item types.

In both organisations, constructed response tasks are trialled and then the output is analysed to ensure that prompts aren't eliciting sensitive responses and that topics are accessible. For instance, in *IELTS* speaking, a prompt about a fictional detective/crime story mostly elicited responses about real-life crime stories that the candidates had read about, so the task had to be rewritten. At CaMLA, a writing prompt about travelling between countries unintentionally elicited strong anti-immigration opinions and had to be revised.

Test assembly

When items, be they constructed response or multiple choice, have successfully passed through all stages of development, including trialling, they are assembled into tests. Both organisations have procedures in place to ensure that all live tests are fair and bias-free.

At Cambridge English Language Assessment, there is a test construction meeting, in which participants check that 'a range of topics/tasks is maintained . . . bearing in mind the range of cultural perspectives desirable', that there is a balance of genders represented, and that there is no overlap in content (Cambridge English Language Assessment 2013b: 49). After this, the Chair and two other external consultants review the test for content, and internal and external proofing is conducted, before the Assessment Manager responsible for the test does a final review and approves it for print.

At CaMLA, tests are assembled by trained Fairness Reviewers in the Assessment Group, who aim to create tests that have a balance of gender in positions of authority, as well as a variety of topics that test takers of various ages, genders, and cultures can relate to. The final assembled test is reviewed by at least two trained reviewers, and an external copy edit is conducted before the test is signed off by the Assessment Manager or Specialist responsible for test compilation.

Conclusion

As Kunnan has argued, fairness 'has to be present at all stages of test development: design, development, piloting and administration and use' (Kunnan 2004:39). In this article, we have seen that fairness is considered throughout the entire life of an item at both CaMLA and Cambridge English Language Assessment.

Though Cambridge English Language Assessment and CaMLA have unique histories that have led them to develop differing methods of safeguarding fairness principles, they share an underlying goal of reducing construct-irrelevance in their tests by avoiding unfair or biased items. Both organisations now have the opportunity and responsibility to learn from each other through increased communication and collaboration when deciding best practices. Just as fairness principles must continuously be revisited and revised, so too should the current practices of each organisation in order to ensure that test takers continue to have the opportunity to take fair and unbiased tests.

References

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (1999) *Standards for Educational and Psychological Testing*, Washington, DC: American Educational Research Association.

Cambridge English Language Assessment (2013a) *Principles of Good Practice: Quality Management and Validation in Language Assessment,* Cambridge: UCLES, available online: www.cambridgeenglish.org/research-and-validation/quality-and-accountability

Cambridge English Language Assessment (2013b) *Work Instructions for Routine Test Production,* Cambridge: UCLES, internal document.

Cambridge ESOL (2006) *Item Writer Guidelines: Information Common to All Papers, FCE, CAE and CPE,* Cambridge: UCLES, internal document.

Cambridge ESOL (2007) *Item Writer Guidelines: Information Common to All Papers, ILEC,* Cambridge: UCLES, internal document.

Cambridge ESOL (2010) *Specifications for Lower Main Suite (LMS) Artwork*, Cambridge: UCLES, internal document.

Hamp-Lyons, L (2000) Fairnesses in language testing, in Geranpayeh, A and Taylor L (Eds) *Fairness and Validation in Language Assessment,* Studies in Language Testing volume 9, Cambridge: UCLES/Cambridge University Press, 30–34.

Kunnan, A J (2004) Test fairness, in Milanovic, M and Weir, C (Eds) *European Language Testing In A Global Context: Proceedings of the ALTE Barcelona Conference July 2001*, Studies in Language Testing volume 18, Cambridge: UCLES/Cambridge University Press, 27–48.

McNamara, T and Roever, C (2006) *Language Testing: the Social Dimension*, Oxford: Wiley-Blackwell.

Investigating grammatical knowledge at the advanced level

FABIANA MACMILLAN TRAINING DEPARTMENT, QUANTA, USA
DANIEL WALTER RESEARCH GROUP, CAMBRIDGE MICHIGAN LANGUAGE ASSESSMENTS, USA
JESSICA O'BOYLE ASSESSMENT GROUP, CAMBRIDGE MICHIGAN LANGUAGE ASSESSMENTS, USA

Introduction

Cambridge Michigan Language Assessments (CaMLA) regularly monitors item distribution to ensure the continued representativeness of the skills tested across test forms. In 2010, as part of its regular test development and revision process, CaMLA conducted a review of the development of grammar items for its Common European Framework of Reference (CEFR) (Council of Europe 2001) C2 level English as a Foreign Language (EFL) certificate test, the Examination for the Certificate of Proficiency in English (ECPE™). Specifically, this review focused on how well the obtained difficulty of recently trialled grammar items reflected developers' expectations. Statistical analyses on item performance revealed that a significant number of trial items tagged as difficult based on test developers' intuition, while successfully discriminating between target-level and lower-level learners, did not prove to be as challenging as anticipated, and thus had to be re-tagged as medium or easy as appropriate. These results are in line with those obtained in several studies, particularly in the area of EFL reading comprehension, which established that expert judges find it difficult to determine how difficult an item will be for test takers (Alderson 1990, Alderson and Lukmani 1989, Lunzer, Waite and Dolan 1979, Sydorenko 2011).

Although the ECPE writers' limited success in predicting the difficulty of grammar items did not represent a decrease in the number of successful trials, it might make it increasingly more challenging to replenish the subset of difficult scorable items in item banks. In order to address this issue, a project was started that aimed to develop a systematic approach to tagging trial grammar items for expected difficulty – one that draws on different resources, rather than rely exclusively on writers' intuition. This article will describe how a theoretically based framework was designed by combining information about frequency of use of individual grammar structures based on a large corpus (Biber, Johanson, Leech, Conrad and Finegan 1999) and the inherent complexity of the same structures from a linguistic perspective (Celce-Murcia and Larsen-Freeman 1999). Further, it will discuss how empirical data on a sample of grammar items was used to inform the design of an additive rubric to infer the combined difficulty of two or more grammatical structures targeted in a single test item. Finally, the paper will consider the limitations of this framework and future avenues for investigation and improvement.

Inferring the difficulty of grammar items during test development

The grammar section in the ECPE assesses the test takers' ability to understand meaning by recognising well-formed phrases and sentences in accordance with the principles governing the assembly of elements of standard American English into meaningful strings (Council of Europe 2001:113). All items are presented in multiple-choice format, each featuring a sentence with one gap followed by four options, one of which is the key.

Tagging rubric

The first step in creating a systematic framework for inferring the expected difficulty of items in development was to standardise the tagging of individual grammatical features as easy, medium or difficult by reference to the proficiency level of the target population (C2 level). In this document, *grammatical features* refer to 'anything that recurs in texts that can be given a linguistic description' (Biber et al 1999:5). These include, for instance, word classes (e.g. relative pronoun), phrasal and clausal categories (e.g. conditional clauses), and other structural distinctions (e.g. inversion). The tagging rubric was informed by second language acquisition (SLA) research, particularly on factors affecting the complexity of specific grammatical structures, as well as the order of acquisition of given grammatical morphemes (Collins, Trofimovich, White, Cardoso and Horst 2009, DeKeyser 2005, Ellis 1990, 2008, Kwon 2005, Larsen-Freeman 1975). The assumption is that the ability to 'maintain consistent grammatical control of complex language', which is characteristic of C2 level learners (Council of Europe 2001:114), entails familiarity with grammatical features at the more difficult end of the complexity spectrum, which tend to be 'late-acquired', or to take time to be mastered (Collins et al 2009). This assumption is in keeping with the 'criterial feature concept' established by the English Profile Programme (Hawkins and Filipović 2012), where criterial features represent certain linguistic properties that are both typical and indicative of second language proficiency at each of the CEFR (Council of Europe 2001) bands.

Four broad perspectives on what can be regarded as a complex, or difficult, grammatical feature may be identified in the SLA literature, namely the acquisition, the pedagogical, the psycholinguistic, and the linguistic perspectives. According to the acquisition perspective, difficult structures are 'those for which the full range of formal/functional aspects develop in stages over time' (Collins et al 2009:337). A strong limitation

of this perspective, however, is that to date very few features of language have been reliably identified as late-acquired forms. The pedagogical perspective focuses on the rule involved in describing the target feature to second language (L2) learners. A feature is considered difficult when it requires a complex explanation, which is challenging to understand and apply. A crucial problem with this view is that it does not account for the fact that simple and accessible rules can take considerable time to master. In contrast, the psycholinguistic perspective sees complexity as a product of the nature and extent of learners' experience with the target language. According to this view, difficult grammatical features will typically be those that occur less frequently in the input (Goldschneider and DeKeyser 2001). Finally, the linguistic perspective uses the inherent characteristics of the target language as a point of departure for determining difficulty. This approach identifies as difficult language features that differ considerably from related features in learners' first language. Celce-Murcia and Larsen-Freeman (1999) have suggested that syntactic complexity also results from the number of transformations required to arrive at the target form. To illustrate, according to this view, the *wh*-question as a prepositional object, which involves seven transformations, is more complex than the simple past tense in English, which involves just one (Collins et al 2009:339). The latter two perspectives – psycholinguistic and linguistic – guided the development of the tagging rubric for ECPE grammar items in 2010. Work done in the English Profile Programme (Hawkins and Filipović 2012) has since greatly contributed to fill gaps in the acquisition and pedagogical perspectives on what aspects of lexico-grammar can be considered advanced, or late-acquired. Following the analysis of a large corpus of learner language, the Cambridge Learning Corpus (CLC) illustrative descriptors were used to distinguish features within each band, making it possible for examiners to accurately differentiate one level from another. The concluding section of this paper will mention how these new developments might help address some of the limitations of the approach taken here.

Grammatical features for the development of the ECPE tagging rubric have been sampled from two main texts: the *Longman Grammar of Spoken and Written English* (Biber et al 1999), which provides information about the frequency at which specific features occur in a 40-million-word corpus, and *The Grammar Book* (Celce-Murcia and Larsen-Freeman 1999:8), which sequences structures 'in an order corresponding to their increasing complexity' from a linguistic perspective. For the sake of clarity, grammatical features are divided into three broad categories: word and phrase grammar, clause grammar, and syntactic choices. These broad categories have in turn been distributed into three bands according to their expected level of difficulty: easy, medium and difficult – these categories referring to the likelihood a C2 level learner will be able to key items correctly.

Features were categorised as easy, medium, or difficult according to their inherent complexity in the sequence provided in Celce-Murcia and Larsen-Freeman (1999) and their frequency distribution as informed by Biber et al (1999). Evaluative comments offered by Biber et al (1999), such as 'common' and 'relatively rare', were used to guide the interpretation of frequency distributions of individual grammatical features and aid in the categorisation of these features for expected difficulty on the ECPE. This accounts for the fact that number of occurrences per million words may be a misleading parameter if considered in isolation. For example, Biber et al (1999:926) describe 'inversion', with over 1,000 occurrences per million words in their fiction and news sub-corpora, as 'relatively rare'. However, comparative constructions of the adjective type *er* + *than*, also occurring just over 1,000 times per million words, including both the American English and the British English sub-corpora, are described as 'common in all registers' (Biber et al 1999:529). Biber et al (1999:40) explain that 'findings reported as percentages are intended to answer research questions relating to proportional use, rather than questions relating to how common a feature is in absolute terms'. Following the same approach, judgements regarding the expected difficulty of grammatical features on the ECPE consider the frequency of individual features relative to other like features.

Features that were described as less complex in Celce-Murcia and Larsen-Freeman (1999) and common or frequent in Biber et al (1999) were categorised as easy (i.e. most learners at the C2 level of proficiency can be expected to successfully key an item focusing mainly on such features); features described as more complex and (relatively) rare were categorised as difficult. Features that were not polarised at either end of the complexity/frequency spectrum were categorised as medium. Features that were described as less complex but whose frequency distribution was low in multiple registers or concentrated mainly in the academic sub-corpus were categorised as medium. No instances of features described as both complex and highly frequent were found. Finally, features that were described as the most basic in terms of complexity and were also extremely frequent in the base corpus (e.g. morphological differences between lexical words) were excluded from the framework as this evidence suggested that items based on these features would be too easy for C2 level learners. All classification decisions were thus a product of each feature's complexity and frequency relative to the proficiency level of the target population. These classifications would have to be adjusted before they could be applied to a test aimed at a lower level than C2.

When tagging items, keys were taken as determining the focus of each item. Therefore, items were tagged as testing the grammatical features appearing in the key alone; features appearing in the stem or distractors but not in the key were disregarded, as in the following example from an ECPE sample test (University of Michigan English Language Institute 2010:11):

> No sooner _____ walked into the office than she was overwhelmed with questions.
>
> a. as she
> b. than she
> c. had she
> d. would she have

Here the key (*had she*) features two grammatical features (perfective aspect and inversion). This item was thus tagged with the two features. The distractors will, no doubt, tap other grammatical features. In the case of the example above, the distractors tap knowledge of subordinating connectors (*as*)

and modals (*would*). However, a decision was made not to code the item for these features. Furthermore, items were tagged for those features in the key that disambiguated it from the distractors. Any features in the key that also appeared in all distractors were not included in the tagging. For example, in the example above, if all of the distractors featured perfective aspect, the item would be tagged as testing inversion only. The tagging process, thus, reflects the item writer's *intended* focus. This decision was made for practical reasons. Banerjee (1995) argues that test takers can arrive at a correct answer to a grammar question by more than one route. Her analysis suggests that there are a number of steps that test takers need to successfully navigate in order to get an item correct. As a result, when test takers get an item wrong, it is not possible to precisely and confidently identify the source of the error. Intended focus was therefore considered to be a more objective way to code items for grammatical features tested – one that avoided second-guessing how examinees might interact with the items.

Additive rubric

One important feature of items testing grammar knowledge at the advanced level is that they typically focus on more than one grammatical feature. This is necessary to ensure that items are challenging enough to provide information differentiating C2 candidates from lower level learners. For this reason, the tagging rubric described in the previous section had to be combined with an *additive rubric*. The purpose of this additive rubric was to offer a systematic means to infer the likely combined difficulty of two or more grammatical features.

In order to design the additive rubric, 62 ECPE grammar items selected from a recent test administration were coded using the new tagging rubric. The items were grouped in three bands – easy, medium, and difficult – based on their obtained difficulty. Obtained difficulties are reported in item response theory (IRT) difficulty values in Table 1. Because IRT difficulty thresholds are assigned to a normal curve, these difficulty bands were chosen to split the items into three groups with the medium difficulty band containing a slightly higher percentage of items than the easy and difficult groups. Items below −2 and above 2 were deemed very easy and very difficult, respectively, and were disregarded because they provide less information about the average test taker due to their distance from the mean.

Table 1: Difficulty bands

Obtained difficulty	Difficulty band
−1.99 to −0.50	Easy
−0.49 to +0.49	Medium
+0.50 to +1.99	Difficult

The difficulty of each grammatical feature tested in each item was then checked against the obtained difficulty band for that item. The resulting additive rubric (Table 2) shows the most frequent combinations yielding each obtained difficulty band:

Table 2: Additive rubric

Expected difficulty band	Feature combinations by difficulty band
Easy	(0+) Easy **+** (1) Medium
Medium	(0+) Easy **+** (2) Medium or (1) Difficult
Difficult	(1+) Difficult **+** (1+) Easy/Medium

The figures within brackets refer to the number of grammatical features tested in an item that fall under the difficulty band indicated on the right. For example, an item will be categorised as easy for expected difficulty if it tests any number of easy grammatical features plus at least one medium feature, as indicated in the tagging rubric. An item will be categorised as medium if its key includes at least two medium features or at least one difficult feature, whether or not it also includes easy features. Finally, an item will be categorised as difficult if it focuses on one or more difficult grammatical features plus one or more easy or medium features, as in the following example taken from an ECPE sample test (Cambridge Michigan Language Assessments 2012:8):

Many people worry about _____ their old age.

a. caring for during
b. to care for
c. being cared for by
d. being cared for in

The key, *being cared for in*, includes three features: passive voice (easy), *-ing* clause (difficult), and preposition (easy). Note that even though the key has two prepositions, this feature was counted only once.

In this small sample, items focusing on a single grammatical feature tended to yield obtained difficulties that were lower than the feature's level of complexity as shown in the tagging rubric. In order to account for that, a rule was introduced where trials testing only one grammatical feature should have their expected difficulty set to one level below that represented by the feature. For example, if an item tests a difficult grammatical feature in isolation, the expected difficulty of that item should be medium.

The next step was to investigate whether the patterns observed in this small sample could be replicated with a different and larger sample and, more importantly, whether expected difficulties based on this framework would successfully match items' obtained difficulty.

Investigating the predictive ability of the framework

In order to test the predictive ability of the newly designed grammar framework, 272 ECPE grammar items were selected from several different administrations and tagged for expected difficulty by three raters. The raters were asked to work individually and indicate what grammatical features each item focused on and use the additive rubric to assign the expected overall difficulty of the item as easy, medium or difficult. Ratings were then analysed for a) rater agreement, and b)

the relationship between expected difficulty and obtained difficulty.

The level of rater agreement was analysed by comparing the overall expected difficulty each rater assigned to each item on a 4-point scale from very easy to difficult. The scale was based on the additive rubric (see Table 2); items were tagged as very easy when all features tested fell under the easy band. A log was kept of the percentage of exact agreement among all raters and between pairs of raters (raters 1 and 2, raters 1 and 3, and raters 2 and 3). Checking group matches against pair matches allowed us to examine how well calibrated the raters were and determine, for example, whether the values obtained through any one rater might be skewing overall counts. In addition to exact matches, adjacent matches were also recorded, where one rater tagged an item's expected difficulty as one band level immediately above or below that selected by the other rater(s).

Results of the initial rater agreement analysis suggest that the raters were not sufficiently calibrated given the low percentage of exact matches. However, Table 3 shows a very high level of rater agreement when adjacent matches are included. Although it would seem that these results are less than impressive for a 4-point scale, it should be borne in mind that the predicted difficulties were not coded directly. Exact matches represented perfect agreement on the combination of individual features tested (and their difficulty) from 46 choices in the tagging rubric.

Table 3: Rater agreement

Type of match	Raters 1 and 2	Raters 1 and 3	Raters 2 and 3
Exact	0.65	0.57	0.50
Exact + adjacent	0.86	0.91	0.83

One factor contributing to the significantly higher number of adjacent matches compared to exact matches is the subjective nature of certain grammatical features. One notable example is idiomatic phrase, which given its low frequency (Biber et al 1999, Moon 1998) and high inherent complexity (Celce-Murcia and Larsen-Freeman 1999) is tagged as difficult. There did not seem to be a consensus among raters as to what could be considered an idiomatic phrase. Out of the 24 occurrences of this feature in the data, only six were tagged as idiomatic phrases by all raters. For instance, the phrase *for the most part* was categorised as an idiomatic phrase by two out of the three raters. Whereas, the phrase *from time to time* was tagged as idiomatic by all raters. This is not unusual, as research in SLA has shown little consensus in terms of what makes a given string of words inherently idiomatic or formulaic (Wray 2002). In most cases, this discrepancy resulted in the overall expected difficulty of the items to be categorised as medium, rather than difficult, in those instances where phrases were not tagged as idiomatic.

The analysis of rater agreement was followed by an examination of the relationship between expected difficulty and obtained difficulty. As previously mentioned, expected difficulty was reported on a 4-point scale, including very easy, easy, medium, and difficult. Items for which raters did not reach a consensus in terms of expected difficulty were assigned a difficulty band based on a) an exact match between two raters, and b) an approximate average of the three ratings (i.e. where each rater gave the item a different rating, the final rating was set to medium). Obtained difficulty refers to the IRT difficulty value obtained for each item.

In order to determine the effectiveness of the classification scheme, we performed a regression analysis using the open-source statistical software environment R. The predicted difficulty categories were classified as dummy variables, which allowed for tests of statistical significance for the difference between means for the categories. However, the use of dummy variables necessitates some subtlety of interpretation. In this analysis, the medium predicted difficulty was used as the intercept in all three regression studies. This means that the other coefficients in the 'Estimate' column are comparisons to the average difficulty obtained for items with a medium predicted difficulty. Therefore, the average obtained difficulty for the very easy tagged items in Table 4 is −1.0545 + 0.6495 = −0.405.

Table 4: Obtained vs. expected difficulty (4-point scale)

Coefficients	Estimate	Standard error	t value	Pr(>|t|)
Medium	−1.0545	0.2946	−3.579	0.000409***
Very easy	0.6495	0.4166	1.559	0.120216
Easy	0.1674	0.3488	0.48	0.631573
Difficult	0.6392	0.527	1.213	0.2226257
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 2.185 on 268 degrees of freedom				
Multiple R-squared: 0.01294, Adjusted R-squared: 0.001892				
F-statistic: 1.171 on 3 and 268 DF, p-value: 0.3211				

Results of the first regression analysis show that, compared to the medium difficulty tag, the very easy and difficult item tags had much higher average IRT difficulties. The easy items also tested more difficult than the predicted medium items, but much less so. In addition, the R-squared and adjusted R-squared tests of overall model fit indicate that very little of the variation found in the data is explained by the predicted difficulties. It should be noted that the very easy tag was used to designate those items that fell outside the scope of the framework, i.e. those items where all grammatical features tested were categorised as easy in the tagging rubric (and therefore did not qualify as any of the expected difficulty bands outlined in the additive rubric) or features that did not occur in the tagging rubric at all. This first analysis confirmed that items under this category do not fit the model. A second regression analysis excluding all items categorised as very easy yielded very similar results (shown in Table 5) for the remaining categories. Since items that were not tagged identically by all raters were marked as medium, it could be that this contributed to the unexpectedly low obtained difficulties for the medium tagged items. As a result, only items that were tagged in complete consensus were used for the second analysis.

Results of this final analysis, albeit with a smaller sample (N = 127), still show a higher average obtained difficulty for easy items than medium items, but to a lesser degree than in the previous analysis. Items that are predicted to be difficult, on the other hand, are on average more difficult than both easy tagged items and medium tagged items. The very low R-squared and adjusted R-squared values indicate poor model

Table 5: Obtained vs. expected difficulty (consensus items only; 3-point scale)

Coefficients	Estimate	Standard error	t value	Pr(>|t|)
Medium	−1.0545	0.3058	−3.449	0.000679***
Easy	0.1674	0.362	0.463	0.644168
Difficult	0.6392	0.547	1.169	0.243883
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 2.301 on 124 degrees of freedom				
Multiple R-squared: 0.02575, Adjusted R-squared: 0.01004				
F-statistic: 1.639 on 2 and 124 DF, p-value: 0.1984				

fit, and the differences between easy-medium and medium-difficult are not statistically significant. Since significance is impacted by sample size, it is probable that statistically significant differences would be found using a larger dataset. Overall, the results do not conclusively show that this framework is effective at predicting IRT difficulty values. However, this analysis was restricted to features tested in the key, excluding features like vocabulary level, domain, sentence length, subject matter, and use of colloquial language. With those features absent from the model, only a certain amount of model fit could be expected.

As a corroborative method of examining success rates, the same data analysed in Table 5 was examined in the form of a contingency table in Table 6 below. The rows represent predicted difficulties, and the columns represent obtained difficulties. The three cells that represent correct predictions, that is, where the predicted and obtained difficulties are the same, are highlighted.

Table 6: Obtained vs. expected difficulty contingency table (consensus items only; 3-point scale)

	Very easy	Easy	Medium	Difficult	Very difficult	Total
Easy	19	**26**	14	13	4	76
Medium	7	12	**9**	9	0	37
Difficult	2	0	6	**3**	3	14
Total	28	38	29	25	7	127

These results are similar to those of the regression, but perhaps more informative. The majority of the items were tagged as easy, and those items tested in all five categories, potentially indicating issues with the additive rubric for difficulty prediction. While a significant number of the items tagged as easy tested difficult, the majority of the items tagged as difficult did test as medium, difficult, or very difficult. Overall, approximately 30% of the items were correctly predicted. Though this is not a majority of the items, it still represents additional information otherwise unavailable to the item writers.

Conclusion

This paper has reported on the development and evaluation of a theoretically based framework for tagging grammar items for expected difficulty. Information about frequency of occurrence of individual grammar structures in a large corpus (Biber et al 1999) and the inherent complexity of the same structures from a linguistic perspective (Celce-Murcia and Larsen-Freeman 1999) have been used to guide the development of a tagging rubric for grammar items written for a C2 level English proficiency test. An additive rubric based on empirical data on item performance was also created to help infer the combined difficulty of different grammatical features tested in a single item. Statistical analyses on a sample of items indicated that expected difficulties based on this framework were somewhat successful at predicting obtained difficulties for the same items.

One limitation of this study was the relatively low level of rater agreement. Further research is needed to investigate whether this was due to rater error (e.g. mistagging grammatical features), or whether adjustments need to be made to the tagging rubric and/or to the additive rubric. As regards the tagging rubric, changes may be needed to account for issues such as subjectivity (e.g. difficulty determining whether a phrase is idiomatic) or washback, where for example, items focusing on difficult grammatical features result in easy items, possibly due to an emphasis given to practising those features in the classroom. One possible avenue for refining the tagging rubric is to incorporate information from the English Profile Programme (Hawkins and Filipović 2012) about the CEFR levels reported for individual grammatical features. The additive rubric used to convert features into predicted difficulties also warrants further consideration as its rules for determining the expected combined difficulty of multiple grammatical features have a significant impact on final counts for rater agreement. As previously mentioned, the combinations featured in the additive rubric mirror the most frequent combinations associated with obtained difficulties categorised as easy, medium and difficult based on a very small corpus of only 62 ECPE items. An informal follow-up experiment was conducted in which the additive rubric was arbitrarily adjusted to designate items as easy, medium or difficult based exclusively on the highest difficulty among the grammatical features tested. For example, if an item tested three easy features and one medium feature, the item would be tagged as medium. Results of this experiment show a significant increase in the number of items where exact agreement was reached among all raters from 127 to 172. The percentage of match between expected and obtained difficulty, however, remained around 30%.

It is important to note that a total of 97 items out of the 272 in this corpus tested grammatical features that are not included in this framework. Of these 97 items, a total of 85 yielded obtained difficulties that fall outside the acceptable minimum and maximum IRT difficulty values for the C2 level test under investigation (−1.99 and +1.99, respectively). Therefore, although this model shows modest results in predicting item difficulty, its tagging rubric seems to provide useful information in terms of what specific grammatical features should be included in C2 level items. Finally, it is possible that, in addition to the limitations observed in the tagging and additive rubrics in this framework, the nature of C2 level grammar itself may be less amenable to systematic approaches to inferring item difficulty than lower levels of proficiency might be. In terms of future steps, it would be beneficial to investigate whether a revised version of the approach to predicting item difficulty presented here might yield more significant results if applied to tests targeting lower levels of grammatical knowledge.

References

Alderson, J C (1990) Testing reading comprehension skills (Part One), *Reading in a Foreign Language* 6 (2), 425–438.

Alderson, J C and Lukmani, Y (1989) Cognition and reading: cognitive levels as embodied in test questions, *Reading in a Foreign Language* 5 (2), 253–270.

Banerjee, J V (1995) *Can grammar testing be justified? A study of the reliability and validity of a discrete-point grammar test*, unpublished Master's Dissertation, Lancaster University.

Biber, D, Johanson, S, Leech, G, Conrad, S and Finegan, E (1999) *Longman Grammar of Spoken and Written English*, Essex: Pearson Education Limited.

Cambridge Michigan Language Assessments (2012) *ECPE Sample Test*, Ann Arbor: Cambridge Michigan Language Assessments.

Celce-Murcia, M and Larsen-Freeman, D (1999) *The Grammar Book: An ESL/EFL Teacher's Course*, 2nd edition, Boston: Heinle and Heinle Publishers.

Collins, L, Trofimovich, P, White, J, Cardoso, W and Horst, M (2009) Some input on the easy/difficult grammar question: An empirical study, *Modern Language Journal* 99 (3), 336–353.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching and Assessment,* Cambridge: Cambridge University Press.

DeKeyser, R M (2005) What makes learning second-language grammar difficult? A review of issues, *Language Learning* 55, 1–25.

Ellis, R (1990) *Instructed Second Language Acquisition*, Oxford: Blackwell.

Ellis, R (2008) Investigating grammatical difficulty in second language learning: Implications for second language acquisition research and language testing, *International Journal of Applied Linguistics* 18 (1), 4–22.

Goldschneider, J M and DeKeyser, R M (2001) Explaining the 'natural order of L2 morpheme acquisition' in English: A meta-analysis of multiple determinants, *Language Learning* 51, 1–50.

Hawkins, J A and Filipović, L (2012) *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework*, English Profiles Studies volume 1, Cambridge: UCLES/Cambridge University Press.

Kwon, E (2005) The 'natural order' of morpheme acquisition: A historical survey and discussion of three putative determinants, *Working Papers in TESOL and Applied Linguistics* 5 (1), 1–21.

Larsen-Freeman, D (1975) The acquisition of grammatical morphemes by adult ESL students, *TESOL Quarterly* 9, 409–430.

Lunzer, E, Waite, M and Dolan, T (1979) Comprehension and comprehension tests, in Lunzer, E and Gardner, K (Eds) *The Effective Use of Reading*, Portsmouth: Heinemann Educational Books, 37–71.

Moon, R (1998) *Fixed Expressions and Idioms in English. A Corpus-Based Approach*, Oxford: Oxford University Press.

Sydorenko, T (2011) Item writer judgments of item difficulty versus actual item difficulty: A case study, *Language Assessment Quarterly* 8 (1), 34–52.

University of Michigan English Language Institute (2010) *ECPE Sample Test*, Ann Arbor: University of Michigan English Language Institute.

Wray, A (2002) *Formulaic Language and The Lexicon*, Cambridge: Cambridge University Press.

A look into cross-text reading items: Purpose, development and performance

FABIANA MACMILLAN TRAINING DEPARTMENT, QUANTA, USA

MARK CHAPMAN ASSESSMENT GROUP, CAMBRIDGE MICHIGAN LANGUAGE ASSESSMENTS, USA

JILL RACHELE STUCKER ASSESSMENT GROUP, CAMBRIDGE MICHIGAN LANGUAGE ASSESSMENTS, USA

Introduction

Readers' ability to integrate information from multiple sources is a skill that educators recognise as increasingly important. A policy brief prepared for the National Council of Teachers of English (NCTE) in the USA stated that 'twenty-first century students need to gather information from multiple sources, evaluate their reliability, and apply their findings effectively' (Gere 2009:16). In a 2001 survey involving university professors and students across North America, the ability to 'synthesize ideas in a single text and/or across texts' was identified by both faculty and students as one of the most important reading skills demonstrated by more academically successful non-native speakers of English in undergraduate and graduate programs (Rosenfeld, Leung and Oltman 2001:21, 27, 49). This is consistent with literacy studies indicating that the ability to integrate information in lengthy materials or multiple sources is one of the key skills distinguishing highly proficient readers from less proficient ones (Kirsch, Jungeblut, Jenkins and Kolstad 2002).

In 2008, in line with these findings and as part of its regular test development and revision process, Cambridge Michigan Language Assessments (CaMLA), then the testing division of the English Language Institute at the University of Michigan (UM-ELI), introduced a new reading comprehension task type referred to as cross-text items. These items require reading more than one text to be answered correctly. Initially this item type was featured exclusively in the Michigan English Test (MET®). This is a multilevel proficiency test, introduced in 2008, that targets Levels A2–C1 on the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001). Cross-text items were subsequently also included in the May 2013 administration of the Examination for the Certificate of Competence in English (ECCE™), which is aimed at the B2 level.

This article will describe the theoretical basis for designing cross-text items, including the specific reading skills they aim to tap with reference to the CEFR. Furthermore, it will discuss how these items are developed and how they perform

statistically. Finally, the paper will consider the limitations involved in developing cross-text items.

Justification for creating cross-text items

The literature in second language acquisition suggests that one of the traits that distinguishes higher-level readers from those at lower levels of proficiency is the ability to use 'reference sources selectively', or bring together information from different sources, understand the relationship between them, and apply this new information to a given purpose (Council of Europe 2001:69, Jamieson, Jones, Kirsch, Mosenthal and Taylor 2000:70, Khalifa and Weir 2009). Cross-text items aim to tap different aspects of this skill by prompting readers to, among other things, locate ideas in multiple texts, compare and contrast features of, or information in, multiple texts, and draw conclusions from multiple texts. In doing so, cross-text items provide test takers with additional opportunities to demonstrate how they interact with different written genres and perform in various real-life reading activities.

With a view to providing a representative sample of the different types of language use learners can be expected to encounter in the criterion environment[1], a variety of reading activities are featured in the MET and the ECCE, including a) reading correspondence, b) reading for orientation, and c) reading for information and argument. It should be noted that, in addition to these, the CEFR mentions 'reading instructions' (Council of Europe 2001:65) as a separate type of reading activity. However, this paper refers to instructions as text types rather than tasks (Alderson, Figueras, Kuijper, Nold, Takala and Tardieu 2006:13). Samples of instructions are therefore included in the ECCE and the MET at target levels (B2 and C1, respectively) as instances of language use within the different types of reading activities covered in these tests.

Cross-text reading units include three to four texts that are loosely connected. For instance, one text may announce a marketing manager job opportunity at a magazine, another may show correspondence between that magazine's marketing office and a subscriber, a third may be a magazine article about management, and a fourth may show a portion of the 'Questions from Readers' section of the magazine in the first and second texts (see Appendix for an example). Therefore, although all texts are somewhat related, they do not necessarily have the same focus. The link between the texts is intended to add interest and reflect real-life situations in which language users come across different texts that have varying levels of relevancy in relation to a specific purpose, for instance, when researching the answer to a question on the internet. Additionally, by prompting readers to access similar background knowledge, the different texts complement one another in providing cues on which to build schema. Finally, the connection between the passages supports the development of cross-text items, which require readers to process and integrate information from different texts. The section that follows will provide details on how cross-text items are developed and the specific skills they focus on.

Developing cross-text items

Reading units with cross-text items are created with several texts of different genres and purposes sharing a common thread in a similar fashion to the types of options that might be returned after an internet user enters a group of keywords in a search engine. This common thread is typically based upon a relation between statements made in different texts, such as cause–effect, problem–solution, compare–contrast, or condition–consequence (MacMillan 2012). In the example below, the cross-text item is focusing on a comparison relation between a portion of a communication between two co-workers considering ordering a new printer (CopyPro) for the office and an online customer review of that same printer.

> What do the authors of the memo and the review agree on about CopyPro?
>
> a. It should not replace a full-sized machine.
> b. It should not be used in draft mode.
> c. It should be used for photographs only.
> d. It is suitable for both home and business use.

A portion of the memo reads:

> I looked online and found some product reviews. Most of the reviews for the CopyPro have been favorable – in fact, several computing websites have named it their top pick. Even though it's aimed at the home-user market (people who want to print photos, for example), its print speed, scan resolution, and copying capabilities are all things that we would take advantage of here in the office.

In turn, a paragraph in the customer review says:

> CopyPro claims its ink is both water resistant and smudge proof. I tested these claims by putting some color pages under running water; the ink did not run, and when the pages dried, the ink did not come off, even with rough handling, which supports CopyPro's claims. This is important for business users who make mailing labels and are concerned about exposure to the weather, and for home users worried about the durability of the photos they print.

The key (option d) highlights a similarity between statements made about CopyPro in two different texts. Although the authors focus on different advantages of the printer (speed, resolution, and printing capabilities in the memo versus water-resistant ink in the customer review), both comments express the opinion that the printer is appropriate for use both at home and in the office. This question exemplifies an important aspect of cross-text items: it should not be possible to answer them without reading two or more texts. (Additional cross-text items can be found on sample tests available at: www.cambridgemichigan.org/resources.met/support-materials.)

Performance of cross-text items

Since their inception with the first MET administration in 2008, statistical results on cross-text items have been extremely good. Results of a comparison between regular and cross-text items based on the same texts suggest that the latter tend to be more successful in yielding acceptable levels of difficulty after the first trial.

[1] For both the MET and the ECCE, the criterion environment includes four domains of language use – personal, public, occupational and educational settings.

A sample of 200 items, 100 cross-text items and 100 regular items randomly selected from a group of recently trialled MET and ECCE reading units, showed that 99% of the cross-text items yielded facility values that fell within the specifications established for each test, while regular items yielded 88% of the successful items. Regular and cross-text items had similar discrimination results, with 94% successful cross-text trials compared to 93% regular. Table 1, below, provides a summary of facility values and discrimination indexes obtained for the items in the sample.

Table 1: Summary statistics on regular versus cross-text items

Regular item FV	Cross-text item FV	Regular item discrimination	Cross-text item discrimination
Min.: 0.2900	Min.: 0.2400	Min.: 0.1800	Min.: 0.1000
1st Qu.: 0.5300	1st Qu.: 0.4300	1st Qu.: 0.3400	1st Qu.: 0.3500
Median: 0.6650	Median: 0.5450	Median: 0.4350	Median: 0.4350
Mean: 0.6476	Mean: 0.5462	Mean: 0.4331	Mean: 0.4459
3rd Qu.: 0.7900	3rd Qu.: 0.6425	3rd Qu.: 0.5200	3rd Qu.: 0.5300
Max.: 0.9300	Max.: 0.8500	Max.: 0.7600	Max.: 0.8800

Because cross-text items prompt candidates to piece together information found in separate texts, they can be categorised as integrated reading tasks (Enright, Grabe, Koda, Mosenthal, Mulcahy-Ernt and Schedl 2000). These are among the more challenging reading tasks as regards the type of match involved in keying the question. *Type of match*, as defined by Jamieson et al (2000:66), is a variable associated with the level of difficulty of reading comprehension items and 'refers to the processes used to relate information in a question or directive to corresponding information in a text, and to the processes used to select an answer from a range of response options'. Therefore, the fact that cross-text items tend to have comparatively lower facility values than regular ones meets design expectations and contributes to the validity of this item type.

Conclusion

This paper has described cross-text items, a reading comprehension task which requires reading multiple texts to be answered correctly. The rationale and theoretical basis behind the development of this new task type and the specific skills it targets were discussed. These tasks offer test takers further opportunities to demonstrate how they interact with different written genres and perform in various real-life reading activities, such as when researching the answer to a question on the internet. In line with recent research in second language reading (e.g. Kirsch et al 2002), cross-text items focus on traits that distinguish higher-level readers from learners at lower levels of proficiency, including the ability to integrate pieces of information from multiple sources and identify the type of relation they hold with one another.

Statistical analyses on a sample of items indicated that cross-text items were as successful as regular reading comprehension questions developed for the same tests in achieving above-average discrimination results. Results also showed that cross-text items perform slightly better than regular items in returning facility values that are appropriate for the target population. Perhaps because they are designed to focus on a specific and inherently challenging reading activity, cross-text items yield facility values that more closely reflect expected difficulty. It is important to note that these items' success in reaching intended levels of difficulty and discrimination should not be taken as an indication that they could conceivably completely replace regular items. Cross-text items cover a narrow range of reading skills, and thus must be complemented by other item types to result in a test that provides a representative sample of the reading activities learners can be expected to encounter in daily life, or has 'cognitive validity' (Khalifa and Weir 2009).

Although cross-text items have very positive results, one limitation of this task type is that it is challenging to develop, resulting in reading units that are more time consuming to produce. It is difficult to avoid content overlap between questions testing information in individual texts and questions focusing on all texts in the unit. Additionally, because all texts are somewhat connected, it can be challenging to design questions that cannot be answered by reading only one of the texts. In terms of future research, it would be beneficial to investigate the processes by which proficient writers develop cross-text units that are effective and free of overlap.

References

Alderson, C, Figueras, N, Kuijper, H, Nold, G, Takala, S and 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.

Cambridge Michigan Language Assessments (2012) *MET Sample Test A*, Ann Arbor: Cambridge Michigan Language Assessments.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching and Assessment,* Cambridge: Cambridge University Press.

Enright, M, Grabe, W, Koda, K, Mosenthal, P, Mulcahy-Ernt, P and Schedl, M (2000) *TOEFL 2000 Framework: A Working Paper*, TOEFL monograph RM-00-4, Princeton: Educational Testing Service.

Gere, A R (2009) Literacy learning in the 21st century: A policy brief produced by the National Council of Teachers of English, *The Council Chronicle* 18 (3), 14–16.

Jamieson, J, Jones, S, Kirsch, I, Mosenthal, P and Taylor, C (2000) *TOEFL 2000 Framework: A Working Paper*, TOEFL monograph MS-16, Princeton: Educational Testing Service.

Khalifa, H and Weir, C J (2009) *Examining Reading: Research and Practice in Assessing Second Language Reading*, Studies in Language Testing volume 29, Cambridge: UCLES/Cambridge University Press.

Kirsch, I S, Jungeblut, A, Jenkins, L and Kolstad, A (2002) *Adult Literacy in America: A First Look at the Findings of the National Adult Literacy Survey*, Washington, DC: US Department of Education.

MacMillan, F (2012) *Developing Independent Reading Skills: Integrating information from multiple sources*, workshop presented at BRAZ-TESOL, Rio de Janeiro, 16–19 July 2012.

Rosenfeld, M, Leung, S and Oltman, P E (2001) *The Reading, Writing, Speaking and Listening Tasks Important for Academic Success at the Undergraduate and Graduate Levels*, TOEFL monograph MS-21, Princeton: Educational Testing Service.

Appendix

This is a reproduction of one of the reading units featured in MET Sample Test A (Cambridge Michigan Language Assessments 2012:22–23). The unit contains both regular reading items, i.e. questions based on a single text, and cross-text items focusing on more than one text. The questions for this unit are available, along with the full test, in the resources section of the CaMLA website (www.cambridgemichigan.org/resources/met/support-materials).

READING

A

Introducing the New CopyPro

The CopyPro's full-featured scanning, copying, and printing capabilities make it perfect for all your home office needs.

- Print images directly from your camera's memory card. No computer required!
- Scan your photos and print them out in many sizes.
- Replace ink cartridges only as colors run out with the special individual ink cartridge system. Four different color cartridges allow you to replace only the colors needed.
- No need to worry about handling photos or other printed material. CopyPro uses quick-drying, smudgeproof inks.
- Edit and fix photos and images with CopyPro's Instant Photo Expert software.

Call to order yours today!

B

MEMO

Jane,

Last week when we discussed purchasing a new copier, you asked me to look into them and to give you my recommendation. I've looked at about ten different models so far. Here's one that I think will be perfect for our office: CopyPro. It has all the features that we discussed, and it is within the budget you mentioned.

I looked online and found some product reviews. Most of the reviews for the CopyPro have been favorable—in fact, several computing websites have named it their top pick. Even though it's aimed at the home-user market (people who want to print photos, for example), its print speed, scan resolution, and copying capabilities are all things that we would take advantage of here in the office.

Look at the attached product description and let me know what you think. If you like this, I'll be happy to take care of ordering one. If you don't, I'll continue looking at other models.

Alan

C

Regular Reviews: Honest Reviews by Ordinary People

Review of the CopyPro
by Steve Wilson, Philadelphia, PA

I am quite pleased with this machine, and I think it offers tremendous value. One of the things I particularly liked about the CopyPro is that it prints at a normal speed with decent quality, which is unusual for printers in this price category. It has five levels of quality, although the draft mode is not recommended—pages are very light and dotty.

CopyPro claims its ink is both water resistant and smudgeproof. I tested these claims by putting some color pages under running water; the ink did not run, and when the pages dried, the ink did not come off, even with rough handling, which supports CopyPro's claims. This is important for business users who make mailing labels and are concerned about exposure to the weather, and for home users worried about the durability of the photos they print.

The CopyPro comes with four separate ink cartridges, meaning users can replace the colors as they run out. This is convenient, and it is cheaper in the long run than using a single cartridge for all colors that has to be replaced more often.

The CopyPro has two memory card slots that can accommodate most types of camera memory cards. I find this to be very convenient—I can plug in my camera's card and print, without connecting my computer. However, the CopyPro Instant Photo Expert software was disappointing. It has minimal features and is not a replacement for full-featured photo editing software—the software that came with my digital camera is much better. Still, CopyPro Instant Photo Expert does let you resize your photos, rotate them, do basic color correcting, and some other things.

In short, I think this is a good machine, and the low price makes it a good value.

The Examination for the Certificate of Competency in English revision project: Maintaining score meaning

NATALIE NORDBY CHEN ASSESSMENT GROUP, CAMBRIDGE MICHIGAN LANGUAGE ASSESSMENTS, USA
JAYANTI BANERJEE RESEARCH GROUP, CAMBRIDGE MICHIGAN LANGUAGE ASSESSMENTS, USA

Introduction

The Examination for the Certificate of Competency in English (ECCE™) was first introduced in 1994 as a high-intermediate level English as a Foreign Language (EFL) examination certifying the ability of teenagers and adults. It comprises four sections, each testing one language skill: listening, reading, writing, and speaking. The results for each section are reported separately and test takers also receive an overall score of Pass or Fail. The overall score is calculated based on the test takers' performance on the individual sections. Detailed information about the ECCE can be found online: www.cambridgemichigan.org/ecce.

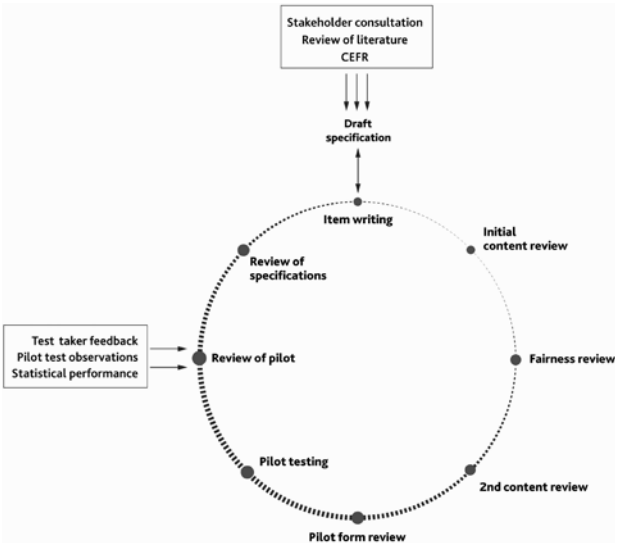
Although the general framework and underlying construct of the ECCE has been maintained, it has, like all CaMLA products, undergone a series of renewal cycles to ensure that it remains an appropriate and effective exam for the test-taking and score-using populations, reflecting the latest research and developments in language learning and assessment (see also, Spillett 2012). Subsequent to the introduction of the Common European Framework of Reference (CEFR) (Council of Europe 2001), its descriptors of language ability have informed task, item, and rating scale development of the ECCE, allowing stakeholders to make more transparent connections between the ECCE and the CEFR descriptors. In 2004, a Claim by Specification was submitted to the Council of Europe demonstrating that ECCE targets the B2 level of the CEFR. A resulting project initiated revisions to the speaking and writing scales which were introduced in 2006 (Dobson, Fleurquin and Ohlrogge 2007). In 2013, revisions to the listening and reading sections were introduced.

The present paper will focus on the updates to the listening section while the new reading item type is discussed in MacMillan, Chapman and Stucker (this issue). A brief outline of the revision process is presented, followed by a description of the main phases in the project including: stakeholder consultation; the main issues uncovered by a review of listening research; and the results of the piloting process.

Overview of the revision process

The revision process began with staff discussions in 2009 and, as discussed in Saville (2003), such a process comprises a number of iterative stages (see Figure 1). During the planning period the project team, led by ECCE Program Manager Sarah Briggs, consulted with stakeholders; re-examined the CEFR scales relevant to the construction of listening tests; and, reviewed the second and foreign language listening literature.

Figure 1: Model of the revision process



The initial specifications for the new item types were prepared in 2010. These were used to develop item sets (testlets) for the first phase of piloting in 2010. Lessons learned from this round of pilot testing – as a result of test taker feedback and pilot test observations as well as item performance – informed additional revisions to the specifications and item sets which were then pilot tested again. The final pilot tests took place in late 2011 after which the test specifications were finalised. A key concern throughout was that the ECCE listening section should maintain or improve its construct coverage and that test takers' scores on the revised May 2013 test should allow the same inferences about their level on the CEFR as comparable scores received by test takers on the tests before the revision.

Stakeholder consultation

Critical to the revision project was feedback from ECCE stakeholders. In February 2010, formal teleconferenced interviews were scheduled with ECCE stakeholders representing major user-groups: language teachers, academic directors and co-ordinators from language schools, test centre administrators, teacher trainers, and applied linguists.

The main aim of the interviews was to explore stakeholder perceptions of the ECCE, helping to highlight areas requiring investigation. A copy of the interview questions was provided to the participants prior to the interviews. Each teleconference began by discussing the target test taker for the ECCE. The stakeholders confirmed that the typical ECCE test taker ranges in age from 13 to 22 years old. Most adults who take the exam do so to enhance their résumés; the ECCE certificate gives test takers access to jobs in the government sector (civil

service) and to professional positions such as hotel personnel, administrative staff, lawyers, accountants, etc. Some test takers also take the ECCE as an entrance requirement for educational programmes such as teacher training.

The second part of each interview focused on the structure of the listening section. Stakeholder feedback was wide-ranging, covering the format of the section, the design of the item types, and the content of input material (stimuli). Table 1, below, summarises the main feedback received.

Table 1: Stakeholder feedback on the ECCE listening section

Feedback category	Comments
Format	• The amount of time provided between questions is sufficient. • Test takers like the multiple-choice format.
Part 1 – Picture listening: item design	• The picture listening items should be retained. As one respondent noted: 'Students are very confident, comfortable with it because of the visual aid . . . it seems easier to them. Adults do not mind the pictures, they like them.'
Part 2 – Radio interview: item design	• The segmented presentation of the interviews helps test takers maintain concentration during an otherwise lengthy stimulus. • Although note-taking is allowed, test takers do not feel they have enough time to write down the information they think will be useful to answering the questions. • Only a single topic is presented on a test, and may not be equally engaging to all test takers. • Test takers perceive this section to be rather difficult.
Content of the input material	• To better engage a broader range of test takers, additional topics are desired.

In mid-2010, students at three different language schools took part in a mini-pilot and were asked for their views of three different types of listening sets: a long radio interview of the type typically included in the ECCE, a short radio interview (similar in format to the existing ECCE radio interview but half the length), and two short talks. The students took each item set under test conditions and then answered a questionnaire. The results of this round of stakeholder consultation indicated that the students found the short talks to be engaging and interesting; they appreciated the range of topics that could be provided. However, they did not have enough time to read the questions prior to the start of the monologue.

In summary, the stakeholder interviews confirmed that the intended purpose of and audience for the ECCE was appropriate and that input materials should be accessible to test takers at or beyond a secondary level of education. Stakeholders were satisfied with the seat-time for the test; therefore, revisions should not alter the length of the listening section, the number of items, or the reading time provided. As important, the picture listening items should be retained and the revision project should focus on the less popular radio interview item type. Short talks emerged as a promising alternative to the radio interview.

Review of theory

With stakeholder feedback in hand the project team turned to the theoretical framework for the revisions. They were mindful of Weir's (2005:18) charge that 'we can never escape from the need to define what is being measured' and also Davies (1984:68) warning that: '. . . in the end no empirical study can improve a test's validity . . . What is most important is the preliminary thinking and the preliminary analysis as to the nature of the language learning we aim to capture.'

This preliminary thinking began with the CEFR (Council of Europe 2001). Perhaps because of its empirical basis (North 2000), the CEFR has been widely adopted both in Europe (for which it was originally developed) and globally. Increasingly, language tests are required to link their reporting scales to the CEFR to allow a comparison of standards across tests and contexts, suggesting that its significance reaches beyond Europe's boundaries. Since the ECCE is aimed at B2 on the CEFR, and it was important to maintain this interpretation of the test scores, the team first analysed the different language contexts in which B2 level users of English are able to understand spoken language, and the strengths and limitations of their listening abilities. The CEFR offers illustrative scales, or Can Do statements, for both overall listening comprehension and specific listening activities. Table 2 shows the B2 descriptors for overall listening comprehension (Council of Europe 2001:66). Some key features have been italicised.

Table 2: B2 descriptors for overall listening comprehension

B2	Overall listening comprehension
	. . . understand *standard* spoken language, live or broadcast, on both *familiar and unfamiliar* topics normally encountered in personal, social, academic or vocational life. Only extreme background noise, inadequate discourse structure and/or *idiomatic usage* influence the ability to understand. . . . understand the main ideas of *propositionally and linguistically complex speech* on both concrete and abstract topics delivered in *standard dialect*, including technical discussions in her/her field of specialisation. . . . follow extended speech and complex lines of argument provided the topic is *reasonably familiar*, and the direction of talk is sign-posted by *explicit markers*.

These descriptors indicate that test takers at the B2 level can follow speech delivered in standard spoken language and can engage with a range of different topics – even when these are abstract or difficult or are presented in complex language – as long as the content is familiar to them, either because it is regularly encountered or because it is in their field of specialisation. Text must be clearly structured and sign-posted and should not be overly colloquial. At this level, test takers are expected to have mastered abilities described under lower levels of competence (A1–B1). Therefore, their listening competencies should go beyond the ability to:

- understand straightforward factual information about common everyday or job-related topics (B1 level)
- understand phrases and expressions related to areas of most immediate priority . . . provided speech is clearly and slowly articulated (A2 level) (Council of Europe 2001:66).

Other descriptors in the CEFR apply these general expectations to four specific listening activities, namely: understanding interaction between native speakers; listening as members of a live audience; listening to announcements and instructions; and, listening to audio media and recordings. With regard to understanding interaction between native speakers, individuals at the B2 level are described as being

able to understand conversations between several speakers as long as the speakers modify their language (i.e. ensure that the utterances are well structured and do not include idiomatic language) (Council of Europe 2001:66). This should go beyond the ability to follow the main points of an extended discussion (B1 level). When listening as a member of a live audience, B2 level listeners 'can follow the essentials of . . . presentation[s] which are propositionally and linguistically complex' (Council of Europe 2001:67), going beyond the need for presentations to be on a familiar topic and also very clearly structured (B1 level). In the case of announcements and instructions, individuals at the B2 level can understand all announcements and messages as long as they are delivered in standard spoken English (Council of Europe 2001:67). They will not be able to understand announcements where the audio is distorted (such as in a sports stadium) or where the information is technically complex (C1 level). Finally, when listening to audio media and recordings, B2 listeners can identify 'speaker viewpoints and attitudes as well as the information content' including 'the speaker's mood, tone, etc.' (Council of Europe 2001:68), as long as the material is delivered in standard dialect.

This close reading of the CEFR was helpful in crystallising the types of challenges that should be presented to test takers, including the opportunity to listen to both conversations (dialogues) and lectures/talks (monologues) on a range of concrete and abstract topics with a degree of linguistic complexity but without overly colloquial language or sound distortions. However, the project team was aware that, as evidenced in several studies (Alderson, Figueras, Kuijper, Nold, Takala, and Tardieu 2006, North 2004, Papageorgiou 2009), the CEFR is a limited resource when it comes to constructing test specifications. Papageorgiou (2009:147) argues that these limitations stem from the fact that the behavioural scales in the CEFR are based on language use in real life, which contrasts with the 'artificiality' of language assessment contexts (Hambleton 2001:100). Therefore, the CEFR was not used as a 'super-specification' (North 2004) for the ECCE listening section but was combined with an extensive review of listening research.

The review resulted in a number of decisions. First, a principled selection was made of the most representative kinds of listening activities second language (L2) users could be expected to encounter in general, academic, and professional settings. A variety of registers ranging along the oral–literate continuum are represented, where one end represents informal, spontaneous speech and the other end represents more formal literate discourse (Tannen 1982).

Second, it was decided that the stimuli should be delivered slightly more slowly than normal American English native-speaker speed, which is 'approximately 188 per minute/3.75 syllables per second' (Griffiths 1992:386). Additionally, the variety of English selected was standard American English and this is used throughout. This decision is somewhat debatable. In the course of their language learning, test takers should become familiar with major native-speaking varieties of English (e.g. British English, Australian English, and American English) and, it can be argued that a language test should offer all these varieties in order, as Taylor (2006:57) argues, to 'reflect varieties of English that enable [test takers] to function in the widest range of international contexts'. Field (2013) also points out that the processing of acoustic cues (of which accent is a part) is an important aspect of the listening construct. However, research looking at the flexibility and adaptability of native-speaking listeners suggests that unfamiliarity with an accent can affect understanding (see Floccia, Goslin, Girard and Konopczynski 2006, Maye, Aslin and Tanenhaus 2008). Additionally the CEFR lists unfamiliar accents among the features that can affect text difficulty (Council of Europe 2001:166). It also states (Council of Europe 2001:27) that even C2 level listeners still need 'time to get familiar with the accent'. With the aim of biasing in the best interests of the test takers, we have taken the position that it is sufficient to test one major native-speaking variety Additionally, in terms of their linguistic characteristics, different national varieties of English (e.g. American English, British English) 'are distinguished primarily by pronunciation differences, and to a lesser extent by lexical and grammatical differences' (Biber, Johanson, Leech, Conrad, and Finegan 1999:17). Therefore, the fact that ECCE listening stimuli feature only standard American English does not affect their representativeness.

The third decision pertained to vocabulary use. Research is currently inconclusive in terms of the relationship between the vocabulary range users of English at the B2 level can be expected to have and the frequency of individual words as informed by native-speaker corpora. In the absence of strong empirical evidence, values suggested by EFL reading comprehension research (e.g. McCarthy 2007, Nation 2001) have been used to guide the choice of vocabulary to be included in the ECCE. In common with other sections of the ECCE, words in the 20–50 per-million frequency range, based on the Corpus of Contemporary American English (COCA) overall count, are considered suitable for testing. Even so, the input may contain words that are lower in frequency than 20 times per million, provided that they are glossed or well supported in context.

Finally, it was decided that the stimuli should contain features that allow the test takers to demonstrate a number of different listening skills (see Buck 2001), including their ability to:

- understand the general idea or main purpose of the stimuli, which includes the main idea or gist and the speaker's general attitude, mood or opinion about the topic
- comprehend specific information, or significant details, in the stimuli, including their understanding of words or phrases in context
- understand implications, or implicitly intended ideas, and draw inferences, draw meaningful conclusions, and/or make predictions
- understand the purpose of specific statements within a given context or proposition.

The new ECCE listening section

The change to the ECCE listening section is presented in Table 3. It shows that Part 1 of the listening section (picture listening items) was retained. The reasons for this were based on both

Table 3: Comparison of the old and new ECCE listening section

Previous ECCE		**Revised ECCE**	
Listening 30 minutes	**Part 1: Picture listening** 30 short dialogues, each followed by a multiple-choice question	**Listening** 30 minutes	**Part 1: Picture listening** 30 short dialogues, each followed by a multiple-choice question
	Part 2: Radio interview A recorded radio interview broken into segments. Each segment followed by several multiple-choice questions, totalling 20 items		**Part 2: Short talks** Four short talks delivered by single speakers on different topics, followed by 4–6 questions each, totalling 20 items

the stakeholder feedback and also the contribution of the picture listening items to the construct of the listening section. The picture listening items (see Appendix 1 for an example) are short conversations, typically no longer than 20 seconds each. They are popular; both younger and older test takers find them accessible. They also represent the oral end of the spoken register spectrum where speech seems unplanned, highly interactive, presupposing an immediate shared context between speakers, and, as such, allows for the use of simple and/or unfinished sentences. They offer good opportunities to test the understanding of social and transactional language use in a variety of contexts.

The changes to the ECCE listening section focused on Part 2. The less popular extended radio interview was replaced by four short talks (see Appendix 2 for an example), each delivered by a single speaker and approximately 1.5 minutes long. This broadened the ECCE listening construct to include monologic speech and to represent both the mid-range and the literate (or written-like) end of the register spectrum. At mid-range, the monologues demonstrate a blend of oral and literate elements; some characteristics of oral communication are still present (e.g. assumed shared knowledge between speaker and audience, redundancies, and pauses) but speech is somewhat planned and sentences are more complex. At the literate end of the continuum, the speaker is detached from the audience, and speech is thoroughly planned or scripted, and linguistically complex, with a minimal amount of redundancy, repetition or pausing (Shohamy and Inbar 1991). The short talks have broadened the content coverage of the ECCE by including a range of topics and offer the opportunity to assess the test takers' ability to understand extended text as members of an audience.

The length of the listening section, including the number of items to be answered, was unchanged. The project team also decided to retain CaMLA's practice of playing the audio once only. This assesses the test takers' ability to process language quickly and automatically, one of the most important of all listening skills (Buck 2001). It is also believed to provide an authentic representation of most real-world listening situations, in which texts are heard only once (Buck 2001:170–171). Although some advocate that playing the recorded stimulus more than once might mimic the effect users experience when asking for clarification in real-life interactions, Buck (2009) points out that these requests are commonly followed by reformulated speech, rather than verbatim repetition. Buck adds that, 'even in the rare case where the same words are repeated, the intonation is always totally different. In fact, listeners never really get the same text twice' (Buck 2009:1).

Piloting

The new ECCE listening section was piloted in two international phases. In October 2010, a pilot test was undertaken in test centres in Europe (N = 366). During this pilot, participants were given a two-part test:

- Part 1: picture listening items
- Part 2: four listening monologue sets, each followed by up to six questions, with stems printed in the test book.

The test takers completed a formal written survey following this pilot test. They were asked about note-taking practices and their response to the printed stems, as well as when/how they interacted with the stems (reading them before, during, or after the stimulus). The results indicated that most test takers did not take notes while listening. The test takers confirmed a desire to preview the questions prior to the start of each talk, though they did not have enough time to read them before the talks began. Without specified time to preview the questions, the listening was less directed and purposeful than desired.

The project team also reviewed the results of the pilot test. In general, the pilot test takers were less able than those who would typically attempt the ECCE. However, even when the pilot test takers' ability was accounted for, the new short talks proved to be more challenging than might be expected for a test taker at the B2 level. As a result, the team refined the items, focusing on those that presented particular difficulties, and revised the test specifications to add preview time.

In October 2011, a second pilot test was conducted in test centres in Europe (N = 318). The structure of the test was the same as that of the first pilot. The test takers were also given a written survey which focused on their views of the short talks. They were asked whether they found the topics interesting; how difficult the items were; and whether there was enough time to respond to the questions. The response rate for the survey was mixed; on average no more than two-thirds of the test takers answered the questions. The results indicated that the majority of the test takers who responded (73%) found the topics of the talks to be interesting; 66% of those who responded found the short talk sets difficult; 46% reported that they did not have enough time to read the questions and answer choices. However, the results of the pilot test were very promising. Though the pilot test takers were (once again) less able than those who would typically attempt the ECCE, when their ability was accounted for, the new short talks proved to be at the right difficulty level for a test taker at the B2 level. At this point, therefore, the item specifications were finalised.

Conclusion

The process of revising the ECCE listening section presented many challenges to the project team. As the test drew close to its 20th birthday, it was important to reflect upon how views 'about the nature of language ability and how it should be taught and tested' had changed (Zeronis and Elliott 2013:22). It was also crucial to maintain the standard of the exam at the B2 level on the CEFR. This narrative has shown how the project team broadened the construct captured by the item types, taking into account stakeholder views as well as the prevailing literature. The renewed ECCE was administered for the first time in May 2013 and the results show that the revised section was as reliable (r = 0.89) as previous ECCE listening sections (see Table 4).

Table 4: Listening section reliability and standard error of measurement (SEM) over time

Administration	Reliability	SEM
May 2011	0.88	0.35
May 2012	0.89	0.33
May 2013	0.89	0.33

Data from post-examination reports indicates that the revised listening section has been well received – a testimony to the consultative approach taken during each phase of the revision project. Additionally, the team's work is not over. CaMLA is already working with stakeholders to introduce an enhanced scoring system for the ECCE in which the speaking and writing sections will be reported on the same numeric scale as the listening and reading sections. The new system will streamline the scoring process and allow more precise information about their performance to be provided to test takers.

References

Alderson, J C, Figueras, N, Kuijper, H, Nold, G, Takala, S and 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.

Biber, D, Johanson, S, Leech, G, Conrad, S and Finegan, E (1999) *Longman Grammar of Spoken and Written English*, Essex: Pearson Education Limited.

Buck, G (2001) *Assessing Listening*, Cambridge: Cambridge University Press.

Buck, G (2009) *Playing the Recording More than Once: Arguments for and Against*, Pre-Conference Workshop at 31st LTRC, Denver, Colorado, USA, March 2009.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.

Davies, A (1984) Validating three tests of English language proficiency, *Language Testing* 1 (1), 50–69.

Dobson, B, Fleurquin, F and Ohlrogge, A (2007) *A Writing Scale Revision Project*, poster presented at LTRC, Barcelona, Spain, June 2007.

Field, J (2013) Cognitive validity, in Geranpayeh, A and Taylor, L (Eds) *Examining Listening: Research and Practice in Assessing Second Language Listening*, Studies in Language Testing volume 35, Cambridge: UCLES/Cambridge University Press, 77–151.

Floccia, C, Goslin, J, Girard, F and Konopczynski, G (2006) Does a regional accent perturb speech processing? *Journal of Experimental Psychology: Human Perception and Performance* 32 (5), 1,276–1,293.

Griffiths, R (1992) Speech rate and listening comprehension: further evidence of the relationship, *TESOL Quarterly* 26 (2), 385–389.

Hambleton, R K (2001) Setting performance standards on educational assessments and criteria for evaluating the process, in Cizek, G J (Ed), *Setting Performance Standards: Concepts, Methods, and Perspectives,* Mahwah: Lawrence Erlbaum, 89–116.

Maye, J, Aslin, R N and Tanenhaus, M K (2008) The weckud wetch of the wast: Lexical adaptation to a novel accent, *Cognitive Science* 32 (3), 543–562.

McCarthy, M (2007) *Assessing Development of Advanced Proficiency Through Learner Corpora*, Center for Advanced Language Proficiency Education and Research, available online: calper.la.psu.edu/downloads/pdfs/CALPER_ALP_Corpus.pdf.

Nation, P (2001) *Learning Vocabulary in Another Language*, Cambridge: Cambridge University Press.

North, B (2000) *The Development of a Common Framework Scale of Language Proficiency*, New York: Peter Lang.

North, B (2004) Europe's framework promotes language discussion, not directives, *Guardian Weekly*, available online: education.guardian.co.uk/tefl/story/0,,1191130,00.html

Papageorgiou, S (2009) *Setting Performance Standards in Europe*, New York: Peter Lang.

Saville, N (2003) The process of test development and revision within UCLES EFL, in Weir, C J and Milanovic, M (Eds) *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002*, Studies in Language Testing volume 15, Cambridge: UCLES/Cambridge University Press, 57–120.

Shohamy, E and Inbar, O (1991) Validation of listening comprehension tests: the effect of text and question type, *Language Testing* 8 (1), 23–40.

Spillett, H (2012) The revision of the Cambridge English: Proficiency writing paper, *Research Notes* 49, 2–5.

Tannen, D (1982) *Spoken and Written Language: Exploring Orality and Literacy*, Norwood: Praeger.

Taylor, L (2006) The changing landscape of English: implications for language assessment, *ELT Journal* 60 (1), 51–60.

Weir, C J (2005) *Language Testing and Validation: An Evidence-based Approach*, Basingstoke: Palgrave Macmillan.

Zeronis, R and Elliott, M (2013) Development and construct of revised Cambridge English: Proficiency, *Research Notes* 51, 22–27.

Appendix 1: ECCE listening Part 1 sample

Script (delivered as audio only)

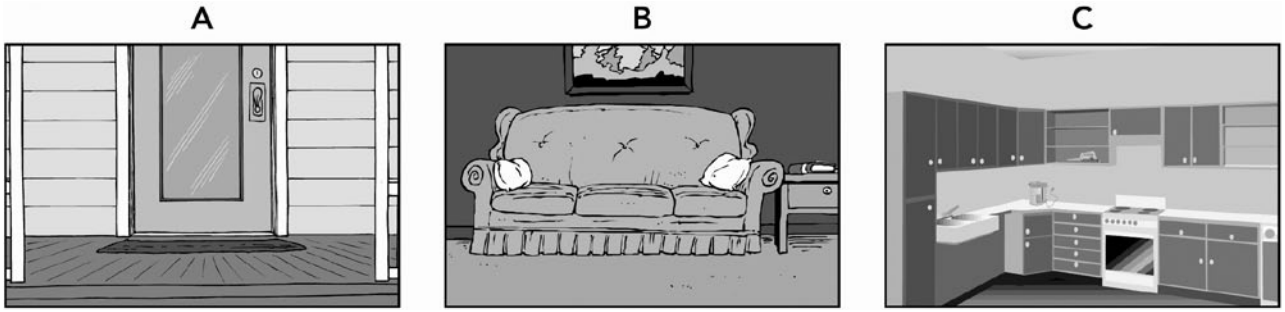
Woman: *Do you remember if we turned off all the lights before we left?*

Man: *I got the kitchen light on my way out, but I don't know about the living room.*

Woman: *No, I got that. And I left the porch light on on purpose.*

Man: *Oh good, I'm glad you remembered that.*

Narrator: *Where is the light on?*



Appendix 2: ECCE listening Part 2 sample

Script (delivered as audio only)

Good afternoon, and welcome to University Radio.

University Art Gallery is hosting a series of photography exhibitions this coming year. The first in the series is called 'Focus on the Garden.' 'Focus on the Garden' opens tomorrow evening. It features over one hundred images by winners of a garden photography contest that was sponsored jointly by University Art Gallery and the Department of Botany. The five winning photographers represent the wide range of exceptional talent in our community. At the exhibition, you will have an opportunity to view their photographs all taken locally in garden settings around the area. These prize winners include, among others, dreamlike images taken when Riverside Gardens was covered with a fine blanket of snow as well as amazing close-ups of such sites as an old apple tree just beginning to blossom in early spring.

The competition was judged by a team of professional photographers led by Susan Cook, founding director of PGP – Professional Garden Photographers. Ms. Cook's photographs often appear on the covers of leading garden magazines including this month's Best Garden Designs.

The exhibition is free and open to the public. For opening night, the winning photographers will be there in person and available to answer questions about their work. For more information about this and future exhibits, please visit the University Art Gallery website at universityart.org.

Questions (delivered as audio and also printed in the test booklet)

Number 31.
What is the purpose of the announcement?

Number 32.
What does the speaker say about the photograph of a tree?

Number 33.
Why does the speaker mention PGP – Professional Garden Photographers?

Number 34.
Why does the speaker mention a website?

Layout and format of response options in test booklet

Questions 31–34
Look at the questions. Then listen to a radio announcement.

31. What is the purpose of the announcement?
to announce winners of a competition
to explain how to win a photography contest
to promote an interest in gardening
to provide information about a coming event

32. What does the speaker say about the photograph of a tree?
It was one of the winning photographs.
It was selected to be in a magazine.
It was taken by a professional photographer.
It was covered lightly with snow.

33. Why does the speaker mention PGP – Professional Garden Photographers?
to encourage listeners to become members
to give the background of a judge
to tell listeners where they can see photographs
to announce whose photos can be seen

34. Why does the speaker mention a website?
The photos can be seen there.
Questions can be submitted there.
More information is available there.
Magazines can be ordered from there.

A discourse variable approach to measuring prompt effect: Does paired task development lead to comparable writing products?

MARK CHAPMAN ASSESSMENT GROUP, CAMBRIDGE MICHIGAN LANGUAGE ASSESSMENTS, USA
CRYSTAL COLLINS ASSESSMENT GROUP, CAMBRIDGE MICHIGAN LANGUAGE ASSESSMENTS, USA
BARBARA ALLORE DAME ASSESSMENT GROUP, CAMBRIDGE MICHIGAN LANGUAGE ASSESSMENTS, USA
HEATHER ELLIOTT ASSESSMENT GROUP, CAMBRIDGE MICHIGAN LANGUAGE ASSESSMENTS, USA

Introduction

Ensuring that high-stakes examinations are fair and consistently produce reliable scores is essential for test designers. However, the process of designing a test to meet these standards can be a difficult one. In particular, certain issues surrounding constructed response items prove challenging.

While problems with scoring speaking and writing tasks have been considerably resolved through advancing constructed response scoring procedures, issues related to prompt equivalence (both within and across test forms) continue to be a concern. One challenging problem for test designers is how to ensure that constructed response items intended to be equivalent are indeed so, offering equal opportunity for test takers to demonstrate their language ability regardless of the prompt they select.

The developers of the writing section of the Examination for the Certificate of Proficiency in English (ECPE™) face this challenge. The ECPE is a standardised advanced-level English as a foreign language examination, aimed at the Common European Framework of Reference (CEFR) (Council of Europe 2001) C2 level, developed and administered by Cambridge Michigan Language Assessments (CaMLA). It consists of compulsory writing, listening, reading, and speaking sections. The ECPE Certificate is recognised in several countries as official documentary evidence of advanced competency in English for academic or professional purposes.

To support the creation of equivalent prompts, both within and across forms of the ECPE, test specifications call for the development of writing prompts in pairs, with certain prompt features carefully controlled. This research investigates whether paired task development generates prompts that elicit comparable writing products and provides test takers equal opportunity to demonstrate their second language writing proficiency.

Review of literature

Prompt variables

The writing assessment literature suggests that the same writer will not always demonstrate a consistent level of performance in response to different writing tasks (Gabrielson, Gordon and Engelhard Jr 1995, Peyton, Staton, Richardson and Wolfram 1990, Smith, Hull, Land Jr, Moore, Ball, Dunham, Hickey and Ruzich 1985). Some writers are better able to write on certain topics and in certain response modes than others and may be favoured by a particular set of prompt variables. While it is generally accepted that different writing tasks can affect test takers' performance, it is less clear which aspects of the prompt cause this variance in quality and/or quantity of response. This section will report on the literature on a range of prompt variables and how they affect responses on writing tests.

Wording of prompt

Different approaches to investigating changes in prompt wording have produced contradictory results. Studies that use differences in holistic scores to examine the effect of changes in prompt wording frequently find there is no statistically significant relationship between the two (Brossell and Ash 1984, Greenberg 1981, Hoetker and Brossell 1989, Leu, Keech, Murphy and Kinzer 1982, Woodworth and Keech 1980). However, other approaches have yielded results that indicate different prompt wording may elicit varying quality and quantity in written responses.

Brossell and Ash (1984) examined whether prompt phrasing had any impact on writing performance. They wrote 21 prompts in two different formats, beginning with an introductory statement followed by either a question or an imperative. Half of the prompts had a personal manner of address and the other half had a neutral manner. Results showed that there was no statistically significant relationship between any of the prompt variables and the assigned holistic scores.

Hirokawa and Swales (1986) went beyond the use of just holistic scores to analyse the effects of different prompt wording. The study asked 32 non-native English-speaking graduate and transfer students at the University of Michigan to respond to two different writing prompts each. The prompts were on the same topic, 'family size', but were presented in both simple and academic versions.

The two sets of responses were analysed syntactically and the following statistically significant differences were identified: 'Simple topic compositions (a) were longer than the academic topic compositions, as measured by both words written per 30 minutes and sentences written per 30 minutes; (b) contained more subordination (per standardised length); (c) exhibited greater use of the first-person, singular pronoun; and (d) contained more morphological errors' (Hirokawa and Swales 1986:344). The differences were significant at the 0.01 level. At the 0.05 level, simple-topic responses contained a lower proportion of Graeco-Latin vocabulary, and had higher

proportions of total errors and syntactic errors. 'Nonsignificant variables included sentence length and frequency of logical connectors, second-person pronouns, and the passive voice.' The authors concluded that the two different prompt types resulted in 'relatively few and relatively small differences' (Hirokawa and Swales 1986:344).

Rhetorical specification

The degree of rhetorical specification in a writing prompt refers to the level of instruction that is provided to the test taker, such as who the intended audience is and the quantity of text that is expected within the response. Brossell (1983) administered three writing prompts which differed in their degree of rhetorical context, to 360 undergraduate education majors at Florida State University. Six essay topics were used and each topic was written at three levels of 'information load' (Brossell 1983:166). Participants were given 45 minutes to complete one essay under test conditions and each response was holistically rated on a four-point scale. The results showed no statistically significant relationships between either topic or level of rhetorical specification and the overall score awarded.

In a similar study, Hoetker and Brossell (1989) systematically varied the degree of rhetorical specification in a writing prompt, using brief or full rhetorical specification and personal or impersonal phrasings. An analysis of the different prompts and responses they generated revealed that there was no statistically significant relationship between rhetorical specification and scores determined by holistic ratings. The authors also concluded that full rhetorical specification had no disadvantageous effect on low-proficiency writers.

Effects of task type

Greenberg (1981) employed holistic scoring to assess whether differences in writing tasks had any effect on writing performance. The tasks in Greenberg's work varied by the cognitive load placed on the test taker. The results of the study showed no statistically significant relationship between tasks with varying levels of cognitive load and holistic scores. Despite the main finding of non-significance, Greenberg also recorded a wide range of discourse variables within the essays produced in her study. The author recorded that 'significant main effects were found in four of the analyses: mean number of T-units, words per clause, and words per essay' (Greenberg 1981:72).

A similar approach was taken by Hoetker and Brossell (1989) and comparable results (no significant differences in holistic scores) were obtained. They designed four different prompts, varied by length and whether the prompt called for a personal or impersonal response. There was much greater variety in the prompt types than those used in Greenberg's study. The four variations employed were brief/personal, brief/impersonal, full/personal, and full/impersonal. The sample size was also much larger than in Greenberg's study. The essays were scored holistically and scores were not directly affected by the amount of rhetorical specification, by the stance of a topic, or by any combination of these features.

The research of Cumming, Kantor, Baba, Erdosy, Keanre and James (2005) into integrated writing tasks for the TOEFL suggests that different task types can elicit different discourse from the same test takers. Participants were required to write six essays each, in response to both independent and integrated tasks. These responses were coded in detail for lexical and syntactic complexity, grammatical accuracy, argument structure, orientation to evidence, and verbatim use of source text. Participants were sorted into three separate proficiency levels and significant differences were found between responses to the independent and integrated tasks. Significant differences in discourse were identified in lexical complexity, syntactic complexity, rhetoric, and pragmatics (Cumming et al 2005:5).

Finally, O'Loughlin and Wigglesworth (2007) examined five different versions of two experimental tasks (one containing 16 pieces of information and the other with 32 pieces of information). A total of 210 English as a Second Language students completed four different tasks (two each at different levels of complexity). The participants were sorted by proficiency and the responses were double rated using both a global band score (a holistic rating) and an analytic scale. The main conclusion drawn was that, 'the results of these quantitative analyses reveal that the differences elicited by the different amounts of information provided in these tasks, and the different types of presentation are very small' (O'Loughlin and Wigglesworth 2007:390).

This finding that holistic and analytic scores varied very little led the researchers to analyse the discourse within the responses to see whether there was any systematic difference in written performance by candidate proficiency level. They looked at task fulfilment, coherence and cohesion, and vocabulary and sentence structure. The discourse analyses revealed that the task with less input (16 pieces of information) produced responses with greater complexity on most measures (structure, organisation, cohesion, subordination, and repetition of key words) across all proficiency levels.

Effects of topic familiarity

The familiarity of the main content of the prompt topic has been reported as an important cognitive variable in determining the difficulty of a writing prompt (Kroll and Reid 1994, Polio and Glew 1996, Powers and Fowles 1998). Test takers write best about 'what is familiar and when the topic taps into their background knowledge (schemata)' (Kroll and Reid 1994:235). Polio and Glew (1996) interviewed 26 English for Academic Purposes (EAP) writers about how they decided which prompt to select on a USA university placement exam. Of the 26 participants, 22 stated that 'they chose or did not choose prompts based on how familiar they were with a topic or how much they had to say about it' (Polio and Glew 1996:42). Topic familiarity was chosen by almost twice as many participants as the reason for selecting a prompt than the second most common reason: the generality or specificity of a prompt.

Summary

The literature is inconclusive in determining the extent to which writing prompts that differ in wording, complexity, and length will affect the responses produced by the same writer in a test situation. There seems to be stronger evidence to support the belief that different tasks (especially integrated versus independent tasks) can cause writers to produce

responses that vary in terms of measurable discourse variables. Differing levels of input material seem to have statistically significant relationships with certain discourse features in written responses. However, there is as yet no consistency in the findings across studies as to which prompt variables influence which features of written language. The literature suggests that minor differences in prompt wording will have little measurable effect on written language. However, variations among prompts, such as length, rhetorical specification, and topic familiarity may influence the writer to produce different quantities and qualities of text.

Materials and method

In order to address the concerns of prompt equivalence outlined above, writing prompts designed for the ECPE are written and developed in matching pairs. The purpose of this approach to prompt design and development is to create writing prompts that are as equivalent as possible and that elicit comparable writing products regardless of the prompt selected by the test taker.

The following pair of writing prompts was identified in order to investigate whether paired writing prompts help to elicit comparable writing products:

> **Prompt 1**: *In the past, farms were usually small, family businesses. However, because of new technology, modern farms tend to be large-scale businesses run by companies. What are some advantages and disadvantages of large-scale farming? Support your ideas with reasons and examples.*
>
> **Prompt 2**: *In some countries, consumers are encouraged to eat mainly food that is grown locally. What are the advantages and disadvantages of eating mainly food that is grown in or near one's own town or community? Support your ideas with reasons and examples.*

The prompts differ in the topic that they are based on but they are paired in terms of:

- the domain they are situated in
- their length, and
- the tasks set for the test taker.

A representative sample of 120 essay responses, 60 for each prompt, was drawn from the November 2012 ECPE test administration (see Table 1 and 2 for sample profile). Overall language proficiency was controlled so that any differences in writing product identified across writing prompts could not be a result of differing language proficiency within the test takers who responded to the prompt. The total number of essays to be selected at each score band was determined by the percentage of test takers in each band for the grammar, cloze, vocabulary, and reading (GCVR) section of this administration (12 A essays, 30 B essays, 40 C essays, 14 D essays, and 24 E essays). This distribution was allocated equally between the first writing prompt topic and the second writing prompt topic (60 essays per prompt).

Table 1: Age of the sample population*

Age		
Age	**# of test takers**	**Percent**
13–16	62	51.67%
17–19	11	9.17%
20–22	11	9.17%
23–25	12	10.00%
26–29	14	11.67%
30–39	8	6.67%
≥ 40	2	1.67%
Total	**120**	**100.00%**

*Percentages do not sum to 100 due to rounding.

Table 2: Gender of the sample population

Gender		
Gender	**# of test takers**	**Percent**
Female	66	55.00%
Male	54	45.00%
Total	**120**	**100.00%**

The 120 essays, which were handwritten during the test, were transcribed into electronic files for analysis. Care was taken to retain all linguistic error from the original responses and the transcriptions were independently spot-checked for accuracy. When the transcribing process was completed, the 120 essays were analysed using a range of discourse variables. The analyses were performed using Coh-Metrix, a program for analysing the complexity of written text from the University of Memphis and the Corpus of Contemporary American English (COCA) (Davies 2008), a monitor corpus of 450 million words from Brigham Young University.

The intention in selecting the variables for analysis was to highlight features which are quantifiable and which operationalise the scoring criteria used in the ECPE writing rating scale. These criteria are fluency, syntactic complexity, lexical sophistication, cohesion, and accuracy. The specific discourse variables selected for the analyses are detailed below.

1. Essay length (word #)
   Essay length is commonly used as a measure of fluency. Development of ideas within a response is difficult to achieve without a certain number of words, and writing assessment research consistently indicates that length of response is one of the best predictors of final score awarded.
2. Average sentence length (ASL)
   Longer sentences tend to require that the writer have command of more complex syntactical rules. Therefore, this was one measure used to assess syntactic complexity.
3. Syntactic left-embeddedness (SYNLE)
   This measure was also used as an indicator of syntactic complexity. In simple independent clauses, the verb tends to occur relatively early in the sentence. A longer delay before the verb occurs is a likely indicator of a more advanced feature such as a dependent clause.
4. Noun phrase density (DRNP)
   Another way to examine syntactic complexity is to look at how modified the language is. The incidence of noun phrases is an indicator of how densely information is packaged within a sentence (Biber and Gray 2010), and the denser the information, the more complex the text is likely to be.

5. Type-token ratio (TTR)
   The type-token ratio reflects the number of unique words in a text divided by the total number of words, and is reported as a decimal between 0 and 1. This variable measures the range of vocabulary in a text. As the type-token ratio is sensitive to the length of the sample analysed, a sample of 190–210 words from each essay was used.
6. Lexical frequency profile (FREQ)
   An analysis was performed on the word frequency of each essay using COCA. COCA divides words into three frequency bands. The first (FREQ1) represents the 500 most common words of English. The third band (FREQ3) represents more infrequent words, beyond the 3,000 most common words of English, and the second (FREQ2) contains all words that fall between the first and second band. If a text has a high percentage of low-frequency vocabulary it is an indication that the text is lexically sophisticated.

   COCA also classifies words by domain and reports a fourth band (FREQAC) whereby words are categorised as 'academic'. This occurs when, based on the texts that comprise the corpus, a word appears in academic texts (such as research journals) at least twice as often as in other types of texts or its overall frequency average. A word tagged as 'academic' is not necessarily a more difficult word, but it does reflect the mode of academic writing.

   Because COCA often does not recognise a misspelled word, it tends to treat the word as highly infrequent (it defaults to the third band). For this reason, before running an essay through COCA, spelling errors were corrected.
7. Latent semantic analysis (LSA) (between sentences and between paragraphs)
   Cohesion is closely tied to the organisation of a text's ideas, so LSA was used to look at semantic similarity between text segments to see how connected the content of the responses were. Coh-Metrix was used to find two measures of LSA; semantic similarity between adjacent sentences (LSAassa) and semantic similarity between adjacent paragraphs (LSAppa).
8. All connectives incidence (CONi)
   Coh-Metrix was also used to look at all connectives to provide insight into connection, organisation and rhetorical control. The value reported is the number of connectives per 1,000 words.
9. Logical connectives incidence
   Logical connectives per 1,000 words were recorded to explore a more specific subset of connectors that are appropriate to argumentative writing.
10. Error count (total number of errors) and errors per 100 words
    Please refer to the section below for a more detailed account of how the error counts were performed. To account for essay length the number of errors per 100 words was also recorded.
11. Total number of spelling errors and spelling errors per 100 words
    The ECPE writing rating scale constitutes the mechanics of writing, which includes spelling errors as part of the scoring criteria. However, the guidelines followed for counting errors did not include spelling. Therefore, spelling errors were separated from the other errors of grammar usage and punctuation.

In order to ensure that error counts were performed consistently, the error count guidelines recommended by Polio (1997:139) were adopted. In an effort to count the number of errors objectively, four experienced CaMLA essay raters read Polio's article and error guidelines and then read five essays independently. The raters then met to clarify some differences, particularly in how errors were counted for collocation and stylistic choices. Following this meeting, each of the four raters independently counted spelling and non-spelling errors in a further 25 essays.

The correlation coefficients for the four raters ranged from a low of 0.862 for non-spelling error counts agreement to a high of 0.984. For spelling error count agreement, the range was from 0.923 to 0.978. The coefficients indicate a high rate of agreement between the four raters in terms of both error counts and spelling error counts. The correlation of 0.862 between two of the raters for non-spelling error counts is notably lower than all other correlations. However, the correlations achieved between the four raters are in line with those reported by Polio (1997:128) of between 0.89 and 0.94. The rater correlations were considered to be sufficiently high to use the error count data in the analyses of the equivalence of the language elicited by the two writing prompts.

Results

To investigate whether there were significant differences between the discourse variables within the responses elicited from the paired prompts, an ANOVA was performed with the prompt as the dependent variable and the discourse variables as the independent variables. A test of homogeneity of variances was run to check that there was equal variance

Table 3: Test of homogeneity of variances

	Levene statistic	df1*	df2*	Sig.
word#	1.259	1	118	0.264
ASL	0.038	1	118	0.847
LSAassa	0.034	1	118	0.854
LSAppa	0.281	1	118	0.597
CONi	0.004	1	118	0.947
CONLOGi	0.118	1	118	0.731
SYNLE	1.440	1	118	0.232
DRNP	0.215	1	118	0.643
TTR	0.234	1	118	0.630
LEXDENS	0.052	1	118	0.820
FREQ1	0.875	1	118	0.352
FREQ2	0.072	1	118	0.789
FREQ3	0.776	1	118	0.380
FREQAC	0.848	1	118	0.359
NonspellERR	1.669	1	118	0.199
NonspellERRper100	2.029	1	118	0.157
SpellERR	0.004	1	117	0.952
SpellERRper100	0.331	1	118	0.566
*df = degrees of freedom				

for each discourse variable and that the assumptions of performing ANOVA were met. Table 3 shows the test of homogeneity of variances results.

None of the discourse variables shows a significant result at the 95% level, indicating that there is no non-standard variance within the dataset and that the assumptions of performing ANOVA are met.

Table 4 shows the results of the ANOVA with the writing prompts as the dependent variable and the discourse variables as the independent variables.

Table 4: ANOVA

		Sum of squares	df	Mean square	F	Sig.
word#	Between groups	5521.633	1	5521.633	1.181	0.279
	Within groups	551860.367	118	4676.783		
	Total	557382.000	119			
ASL	Between groups	40.976	1	40.976	1.595	0.209
	Within groups	3031.392	118	25.690		
	Total	3072.368	119			
LSAassa	Between groups	0.003	1	0.003	0.950	0.332
	Within groups	0.354	118	0.003		
	Total	0.357	119			
LSAppa	Between groups	0.006	1	0.006	0.510	0.476
	Within groups	1.382	118	0.012		
	Total	1.388	119			
CONi	Between groups	741.171	1	741.171	1.840	0.178
	Within groups	47523.959	118	402.745		
	Total	48265.130	119			
CONLOGi	Between groups	5715.919	1	5715.919	24.788	0.000
	Within groups	27209.519	118	230.589		
	Total	32925.438	119			
SYNLE	Between groups	7.247	1	7.247	2.324	0.130
	Within groups	368.048	118	3.119		
	Total	375.295	119			
DRNP	Between groups	669.840	1	669.840	0.648	0.423
	Within groups	122034.059	118	1034.187		
	Total	122703.898	119			
TTR	Between groups	0.000	1	0.000	0.084	0.773
	Within groups	0.300	118	0.003		
	Total	0.301	119			
FREQ1	Between groups	20.833	1	20.833	0.935	0.335
	Within groups	2628.633	118	22.277		
	Total	2649.467	119			
FREQ2	Between groups	24.300	1	24.300	2.662	0.105
	Within groups	1077.167	118	9.129		
	Total	1101.467	119			
FREQ3	Between groups	0.008	1	0.008	0.001	0.975
	Within groups	1007.917	118	8.542		
	Total	1007.925	119			
FREQAC	Between groups	29.008	1	29.008	2.814	0.096
	Within groups	1216.583	118	10.310		
	Total	1245.592	119			
NonspellERR	Between groups	437.008	1	437.008	3.067	0.082
	Within groups	16812.117	118	142.476		
	Total	17249.125	119			
NonspellERRper100	Between groups	89.722	1	89.722	4.250	0.041
	Within groups	2491.019	118	21.110		
	Total	2580.742	119			
SpellERR	Between groups	16.307	1	16.307	1.190	0.278
	Within groups	1603.390	117	13.704		
	Total	1619.697	118			
SpellERRper100	Between groups	4.881	1	4.881	2.732	0.101
	Within groups	210.837	118	1.787		
	Total	215.718	119			

As can be seen from the ANOVA results, the only discourse variables that produced significant differences between means across prompts are the incidence of logical connectors and the number of non-spelling errors per 100 words. The descriptive statistics show that Prompt 1 elicited an average of 50.2 logical connectors per 1,000 words of text. Prompt 2 elicited an average of 64.0 logical connectors per 1,000 words of text and this is a statistically significant difference at the 95% level. The descriptive statistics also show that Prompt 1 elicited 8.3 non-spelling errors per 100 words on average. Prompt 2 elicited an average of 6.6 spelling errors per 100 words. This is a statistically significant difference. The question of whether it is a meaningful difference in the testing context will be discussed in the following section.

There were no significant differences recorded for any of the other discourse variables. This indicates that fluency, syntactic complexity, and lexical sophistication, as operationalised within the existing research design did not vary significantly within the essays written in response to the two different prompts.

Discussion and implications

Each response within the study was analysed in terms of 17 discourse variables. In only two of the 17 was a statistically significant difference found between the mean values when comparing the language elicited by the two prompts. Only one of the four discourse variables that operationalise cohesion yielded a statistically significant difference. Similarly, only one of the four discourse variables that operationalise accuracy yielded a statistically significant difference. While the two statistically significant different discourse variables should not be ignored, the ANOVA data does provide a good indication that the writing prompts studied are eliciting broadly comparable written language.

The two significant differences are not easy to interpret. There is a significant difference in the average number of logical connectors (*and; or*) but the other variables that operationalise cohesion (incidence of all connectors and latent semantic analysis between sentences and paragraphs) do not show any significant differences. Both prompts set the test taker the same tasks: list advantages and disadvantages of an idea and provide supporting reasons and examples. One possible explanation for the difference in the number of logical connectors elicited by the two prompts is the vocabulary required to provide supporting examples. Prompt 2, which elicited significantly more logical connectors, requires writers to produce supporting examples about food. Prompt 1 requires supporting examples based around business scale and agricultural technology. The vocabulary required to write a supporting argument for Prompt 2 is likely more straightforward and easily accessible for the test population. A re-reading of essays across the proficiency bands provides some support for this explanation. Some writers provide strings of supporting examples in response to Prompt 2 connected together using *and/or*:

> *In my hometown for example, we grow oranges and apples but we don't grow bananas.*
>
> *For example, it will not contain chemical substances and the products will always be fresh and good.*

If the supporting examples are based on vocabulary that is easily retrievable from the mental lexicon, it is likely to be easier for test takers to produce several supporting examples, which may be strung together as in the examples provided above. The lexis required to support Prompt 1 is less straightforward and somewhat more specialised, and this may be a partial explanation for the significant difference in logical connectors elicited by the two prompts.

Prompt 1 elicits more error (spelling and non-spelling error) than Prompt 2 in terms of both total number of errors and number of errors per 100 words. The only significant difference though is in the number of non-spelling errors per 100 words. Prompt 1 elicited an average of 8.3 non-spelling errors per 100 words on average while Prompt 2 elicited an average of 6.6 non-spelling errors per 100 words. Although this difference is significant, it is questionable whether the difference is meaningful. Responses to Prompt 2 tend to be longer than those to Prompt 1, and the mean number of errors in responses to Prompt 1 is 22, while for Prompt 2 it is 18. Would raters be influenced by this additional amount of error? Would it be distracting? This would depend on the severity of the error, but these relatively small differences in error quantity may well not be very meaningful in terms of the overall impression created by the responses. Further investigation into error severity and its impact on the reader or rater would be necessary before it would be possible to conclude that the significant mean difference in error quantity is an indicator of meaningful differences in written product.

There are limitations to the extent to which the findings of this work may be generalised. The sample population consisted of candidates from the ECPE, an exam targeted at the C2 level of the CEFR. Whether findings would be similar for responses to a pair of prompts developed for and administered to individuals at a lower level of language proficiency is unclear. Replicating this work with a more diverse sample population would indicate whether the findings are generalisable. Finally, while the sample size is adequate for the scope of this work, it would be of interest to see whether the findings remain non-significant with a larger and more diverse sample. Despite these limitations, the authors do believe that the findings reported here give a clear indication of the merits of controlling prompt variables within prompt design. The implications of this work are discussed further below.

The implication for writing task design within the limitations of the current study is that the paired development and administration of writing prompts appears to be effective. The findings of this study indicate that standardising the domain, length and task wording may make a positive contribution to prompt equivalence, and these findings suggest that paired writing prompts may help elicit comparable writing products. The more test designers can do to standardise their writing prompts without making them overly predictable and formulaic, the better chance there is of these prompts eliciting language that is broadly comparable across test forms and administrations.

Conclusion

The work reported above indicates that there are identifiable features of writing prompts that may be controlled for and that doing so may positively contribute to prompt equivalence. The prompt features controlled for in this work are domain, length and task wording. Controlling for these features in prompts intended to be equivalent appears to elicit language in the writing products that is comparable in terms of fluency, syntactic complexity, lexical sophistication, cohesion and accuracy. The comparability of writing products is an indicator that raters will be presented with texts that may be processed and scored consistently using a well-designed rating scale. Test specifications can be written to control for these prompt variables and such specifications could make a positive contribution to the equivalence of test forms both within and across test administrations.

Gaining a greater understanding of the relationship between prompt variables and measurable linguistic features of written responses will assist in designing writing prompts that are equivalent for test takers in terms of opportunity and difficulty. Understanding how writing prompts can be made equivalent for test takers is a good step on the road to fair and consistent assessments that present an equal opportunity and challenge to test takers regardless of when they choose to take the test.

References

Biber, D and Gray, B (2010) Challenging stereotypes about academic writing: Complexity, elaboration, and explicitness, *Journal of English for Academic Purposes* 9, 2–20.

Brossell, G (1983) Rhetorical specification in essay examination topics, *College English* 45 (2), 165–173.

Brossell, G and Ash, B (1984) An experiment with the wording of essay topics, *College Composition and Communication* 35 (4), 423–425.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching and Assessment,* Cambridge: Cambridge University Press.

Cumming, A, Kantor, R, Baba, K, Erdosy, U, Keanre, E and James, M (2005) Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL, *Assessing Writing* 10 (1), 5–43.

Davies, M (2008) *The Corpus of Contemporary American English: 450 million words, 1990–present*, available online: corpus.byu.edu/coca/

Gabrielson, S, Gordon, B and Engelhard, Jr, G (1995) The effect of task choice on the quality of writing obtained in a statewide assessment, *Applied Measurement in Education* 8 (4), 273–290.

Greenberg, K L (1981) *The Effects of Variations in Essay Questions on the Writing of CUNY Freshmen*, New York: CUNY Instructional Resource Center.

Hirokawa, K and Swales, J (1986) The effects of modifying the formality level of ESL composition questions, *TESOL Quarterly* 20 (2), 343–345.

Hoetker, J and Brossell, G (1989) The effects of systematic variations in essay topics on the writing performance of college freshmen, *College Composition and Communication* 40 (4), 414–421.

Kroll, B and Reid, J (1994) Guidelines for designing writing prompts: clarifications, caveats, and cautions, *Journal of Second Language Writing* 3 (3), 231–255.

Leu, D J, Keech, K L, Murphy, S and Kinzer, C (1982) Effects of two versions of a writing prompt upon holistic score and writing processes, in Gray, J R and Ruth, L P, *Properties of Writing tasks: A Study of Alternative Procedures for Holistic Writing Assessment,* Berkeley: University of California, Graduate School of Education, Bay Area Writing Project, 215–219.

O'Loughlin, K and Wigglesworth, G (2007) Investigating task design in academic writing prompts, in Taylor, L and Falvey, P (Eds) *IELTS Collected Papers: Research in Speaking and Writing Assessment,* Studies in Language Testing volume 19, Cambridge: UCLES/Cambridge University Press, 379–421.

Peyton, J K, Staton, J, Richardson, G and Wolfram, W (1990) The influence of writing task on ESL students' written production, *Research in the Teaching of English* 24 (2), 142–171.

Polio, C (1997) Measures of linguistic accuracy in second language writing research, *Language Learning* 47 (1), 101–143.

Polio, C and Glew, M (1996) ESL writing assessment prompts: How students choose, *Journal of Second Language Writing* 5, 35–49.

Powers, D and Fowles, M (1998) *Test Takers' Judgments About GRE Writing Test Prompts* (GRE Board Research Report No. 94-13R), Princeton: Educational Testing Service.

Smith, W L, Hull, G A, Land Jr, R E, Moore, M T, Ball, C, Dunham, D E, Hickey, L S and Ruzich, C W (1985) Some effects of varying the structure of a topic on college students' writing, *Written Communication* 2 (1), 73–89.

Woodworth, P and Keech, C (1980) *The Write Occasion,* Berkeley: University of California, Graduate School of Education, Bay Area Writing Project.

Nativelike formulaic sequences in office hours: Validating a speaking test for international teaching assistants

ILDIKO PORTER-SZUCS ASSESSMENT GROUP, CAMBRIDGE MICHIGAN LANGUAGE ASSESSMENTS, USA

UMMEHAANY JAMEEL ASSESSMENT GROUP, CAMBRIDGE MICHIGAN LANGUAGE ASSESSMENTS, USA

Introduction

Performance tests attempt to replicate the real-life setting and standards that test takers will face in the context where they will perform their duties (Wiggins 1989). The International Teaching Assistant Speaking Assessment (ITASA™), a performance test, is no exception. Since 1983,

this standardised test has been used at the University of Michigan to assess the English language proficiency of prospective international teaching assistants (ITAs) at the high-intermediate to advanced levels of proficiency. This and other tests followed in the wake of 'a groundswell of [. . .] complaints' on university campuses around the United States about the spoken proficiency of ITAs, which 'led administrators at several colleges to address the 'foreign TA problem' (Bailey 1983:309). Since its inception in the early 1980s, ITASA (under its local name Graduate Student Instructor Oral English Test) has been fine-tuned several times. Feedback from stakeholders, such as students, the departmental faculty, the English language faculty, and administrators has contributed to the test taking on its current shape. The latest version, released in 2013, has since been made available to other institutions.

Conceptualised from the sociocultural and functional perspectives of transactional and interactional language by Brown and Yule (1983), ITASA's design closely mirrors the contexts that teaching assistants (TAs) typically operate in at North American universities. This holistically-scored test consists of four tasks: a warm-up conversation, a lesson presentation, an office-hour role play, and a listening comprehension task of videotaped student questions. It takes 14–15 minutes to administer the latest version of ITASA. It involves one test taker and two to three evaluators: two trained English language specialists (one who plays the role of a student in Task 3 Office-Hour Role Play), and one optional departmental representative. The Office-Hour Role Play, which serves as the object of enquiry for the present study, is 2–3 minutes long. In this semi-structured task, one of the evaluators, who acts as a student of the ITA, comes to office hours to discuss an issue. The issue is a scripted scenario, which the 'student' presents verbatim to the ITA. The scenarios are written by professional test developers. The inspiration for the scenarios is taken from the Michigan Corpus of Academic Spoken English (MICASE), which is a 'collection of transcripts of academic speech events' (Simpson, Briggs, Ovens and Swales 2002). The presentation of the initial scripted scenario is followed by an immediate response from the ITA candidate. The 'student' then asks follow-up questions and the ITA responds to them. The exchange, along with the other three tasks, is rated and one holistic score is given at the end. Table 1 depicts the relevant excerpt of the evaluation criteria. The four-point rating scale ranges from *A (Very Strong Proficiency)* to *D (Inadequate Proficiency)*. At the University of Michigan, the cut score between 'approved to teach' and 'not approved to teach' lies between *B (Effective Proficiency)* and *C (Limited Proficiency)*. The complete list of evaluation criteria and the rating scale are available online: www.cambridgemichigan.org/itasa.

Research over the years has established the test's effectiveness (see Briggs and Hofer 1991, Briggs, Madden and Myers 1994, Plough and Bogart 2008, Plough, Briggs and Van Bonn 2010). In their study, Plough et al examined the lesson, office hour, and video tasks of ITASA. Their results indicate that of the various evaluation criteria, pronunciation and listening comprehension were the most significant predictors of approval for teaching duties. The Plough et al study also looked at transactional and interactional competences in general but could not conclude that these criteria would be statistically significant predictors of passing the test. However, they did find that evaluators made numerous comments about 'lexical range and grammatical errors', which are assessed under both transactional competence and interactional competence. Plough et al hypothesised about these comments that 'given their saliency to evaluators, these features may represent discrete components of the speaking construct' (2010:251–252).

Table 1: Partial evaluation criteria

Transactional competence	Use of grammar, vocabulary, and nativelike formulaic sequences: • to produce organised, coherent explanations • to summarise • to highlight key points • to paraphrase to clarify content • to respond thoroughly to questions Ability to identify and repair misunderstandings of content Ability to use questions to promote discussion and guide instruction
Interactional competence	Comprehension and use of verbal (e.g. questions, back channels) and nonverbal (e.g. eye contact, body posture) strategies to promote interaction and support communication Use of nativelike formulaic sequences to establish and maintain social relationships Use of pragmatically appropriate language (e.g. use of hedges or softeners, when appropriate, to deny a request or disagree) Ability to adjust language style to situation appropriately and to use appropriate register

This study raised an intriguing question about whether lexico-grammatical features of language influenced the outcome of the test. Therefore, in the present study we set out to investigate a feature that spans lexis and grammar: the production of *nativelike*[1] *formulaic sequences* (NFSs) within Task 3, the Office-Hour Role Play. By formulaic sequence, we largely refer to 'a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar' (Wray 2002:9). However, like Ellis, we prefer not to make a firm distinction between formulas that are on the one end of the lexical–grammatical continuum 'heavily entrenched and conventionalised formulaic units' and on the other 'loosely connected but collaborative elements' (2012:25). We examine both types of word strings so long as they appear to be correct, appropriate, and formulaic within the academic speech community. In terms of their function, these

[1] For more information on 'native norms', see Ellis (2012:29).

sequences enable speakers and listeners to communicate fluently (Nattinger and DeCarrico 1992). It is especially when the speaker's attention is focused on the message that formulaic sequences (FSs) serve as energy- and time-saving devices.

As NFSs are widely used both to establish relationships and to express factual and propositional information, the office-hour task of ITASA was chosen as a vehicle for this investigation. This task presents a unique opportunity for the study of formulaic sequences in both interactional and transactional situations. Of the two types of topics that are prevalent in authentic office hours – content and housekeeping – the test is designed to elicit only the latter, which refers to the language used in an academic setting to discuss administrative and classroom matters. The scripted scenarios have a general focus so that prospective ITAs from all departments could engage with them with equal ease.

Two research questions guided our enquiry:

1. To what degree does the volume of NFSs produced in Task 3 Office-Hour Role Play contribute to the final rating?
2. Does Task 3 Office-Hour Role Play elicit NFSs similarly to the real-life setting?

Methodology

The setting for this study was the University of Michigan (UM), Ann Arbor, where ITASA is administered under its local name: the Graduate Student Instructor Oral English Test (GSI OET). Participants were recruited through their departments. They comprised all students who had recently taken (passed or failed) the test and who agreed to being videotaped for the purpose of new examiner training materials. All who responded to the initial recruitment email were accepted into the study. Table 2 depicts the demographic information of the 30 participants, which is similar in composition to that of the test taker population at the UM in a typical year.

Table 2: Participant background

Test taker L1	Male	Female	LSA	Engineering	Other	Total
Chinese	10	5	8	7		15
Farsi	2	1	1	2		3
Hungarian		2		1	1	2
Korean	6	1	1	5	1	7
Polish		1	1			1
Spanish	1	1	1	1		2
Total	19	11	12	16	2	30

Data collection and analysis

In order to answer the first research question, recordings of the Office-Hour Role Play were transcribed and transcriptions were verified by both of the investigators. In the first step, all 30 transcripts were annotated independently by both investigators. The list of NFSs and any discrepancies were then compared. A consensus list of contiguous and discontiguous NFSs was created, and the list was then checked for correctness and appropriacy. Grammatical correctness was determined by the researchers' expert judgement. Appropriateness for the context was viewed from a pragmatics perspective, as the term 'formula' in pragmatics 'emphasises a speech community's preference for a particular string' (Bardovi-Harlig 2012:2). Incorrectly and inappropriately used word strings were discarded. This eliminated the challenge of having to guess at the sequences that were attempted. It also eliminated the problem of setting tolerances for degrees of incorrectness that would still be counted.

Next, NFSs were identified with the help of AntConc (Anthony 2011). All the non-native speakers' (NNSs) transcripts were loaded into the concordancer, which was set to find 2–8 Grams. The upper limit was eventually set at 8 because a higher number did not result in any NFS hits. The lower limit was set at 2 and not higher so as not to miss any possible, correct, and appropriate sequences, however short they may be. As Wray points out, some formulaic sequences may be infrequent but highly entrenched (2002). The minimum N-Gram frequency was kept at 1 for the same reason. This examination yielded a further refined list of NFSs. Post-hoc analysis was also applied to reject any sequences that resulted from the imprecision of the computer tool, such as unintentional NFSs resulting from across sentence boundaries. The researchers also discarded non-formulaic instances of sequences that may have formulaic meanings in some contexts. For instance, 'like what' appeared twice in the 2-Gram search in AntConc. One of these instances was a NFS, 'Umhm. Like what? Dissecting . . .'. The other by another test taker was not: 'So you can know like what is going to be.' Finally, repetitions of a FS due to word searches or dysfluency were only counted once. An example from the reference corpus (explained later) included the following sequence: 'a little bit a little bit a little bit odd.'

The transcripts were then purged of all speech fillers denoting dysfluency, such as 'hmm', 'uh', 'um', and 'ah'. The speech fillers that were kept were those that carried a meaning, such as agreement ('yeah' or 'yep'), disagreement ('nah' or 'uh-uh'), and surprise ('oh' or 'ah'). In preparation for the Corpus of Contemporary American English (COCA) analysis, all contractions in the transcript were separated into two words, so 'can't' became 'ca n't' and 'gonna' became 'gon na'. This step was taken so that contracted forms of words could also be included as formulaic sequences.

In the third step, this list was verified through COCA (Davies 2008). The frequency of every earlier identified NFS was checked within the entire corpus, rather than any given subcorpus. This decision was made to cast as wide a net as possible so as not to miss any legitimate NFSs that were not present in one subcorpus. Unlike the customary minimum 10 occurrences per million, no minimum frequency threshold was set for acceptance into the final list of NFSs. The reason was that COCA is not a corpus of spoken academic language, nor is it likely to contain instances of NFSs specific to the local context. One such example that yielded zero results but was kept as a legitimate formulaic sequence was 'on CTools', as in 'I 'm gon na upload this on CTools' (referring to the web-based course management system used at the UM). Wherever possible, the longest possible string of NFSs was kept together. However, as in the above example, it is arguable whether this is one 8-word formulaic sequence or

the combination of three shorter ones: 'I 'm gon na' + 'upload this' + 'on CTools'. In order to circumvent the problem of how to count FSs, the number of words contained within each sequence was counted. Therefore, whether the example here was one sequence of eight words or three sequences of four, two and two words, respectively, did not make a difference. The final list contained every correct and appropriate string of words that both researchers agreed on after multiple examinations with the help of AntConc and COCA. This list contains clauses, e.g. 'I do n't know'; phrases, e.g. 'a little bit'; contractions pronounced as single words e.g. 'wan na'; and interjections, e.g. 'oh ok'.

The final ratings used for this study were awarded at the time of the live test administrations by the evaluators who participated in each test. Based on the total number of words spoken by each ITA candidate in each transcript and the total number of words contained in NFSs, a one-way ANOVA was run using SPSS.

In order to answer the second research question – whether Task 3 Office-Hour Role Play elicits nativelike formulaic sequences similarly to the real-life setting – a reference corpus had to be found. MICASE was chosen for this purpose because it most closely resembles the base corpus under investigation. In our search for native, American-English-speaking TAs conducting office hours, we initially browsed only office-hour transcripts while filtering for American-English-speaker status and academic position/role. This search yielded seven results. The type of language under investigation in this study is 'housekeeping' language between TAs and their students, which we described earlier as the language used in an academic setting to discuss administrative and classroom matters. This genre, however, can occur in other types of setting as well. Therefore, the search was expanded to discussion sections and lectures. This resulted in 15 transcripts in total. Each transcript was stripped of content-specific language by an assistant and checked by both researchers.

Similarly to the base corpus, in the reference corpus contractions were separated and NFSs were extracted following the steps outlined above. These transcripts were also purged of speech fillers denoting dysfluency. The only difference was that native speaker (NS) corpora were not evaluated for correctness and appropriacy. If it was a recognisable NFS based on expert judgement, AntConc, or COCA, it was included in the data set. A one-way ANOVA with five levels (NS and A, B, C and D ratings for NNSs) was conducted subsequently.

After the ratio of production of NFSs within both data sets was calculated, the transcripts were re-examined for any salient patterns from a qualitative standpoint. It quickly became apparent that the production of contractions was worth a second look. Therefore, all NFSs were re-examined. If contractions were used, the frequency of both the contracted and the equivalent full forms was checked in COCA. First, however, we needed to determine whether the spoken subcorpus or the overall corpus should be used for this comparison. We compared the results obtained from the spoken subcorpus to the overall corpus on a dozen cases to see if any patterns would emerge. However, no perceivable difference was found in the ratio of results, so the spoken subcorpus was abandoned in favour of the overall corpus.

Having established that the frequency counts would be based on COCA overall, the frequency of every contraction within the NFSs was noted and compared against the full forms. For instance, the frequencies of 'ca n't', 'cannot', and 'can not' were compared. Where the test taker used a contracted form that was more common than the full form, a point was assigned. In the rare instance that the chosen contracted form was less frequent than the full form, the use of the contraction was examined for appropriacy for the socio-pragmatic context. Based on the video rather than the transcript alone, we evaluated whether pauses, thought groups, or other considerations may lead us to reject the contraction as an incorrect and inappropriate one. If we failed to reject it, the use of contractions was assigned a point. Subsequently, we looked for missed opportunities within NFSs. We looked up the COCA overall frequencies of sequences that could have been contracted but were not. In cases where a test taker did not use a contraction within an NFS in spite of its higher frequency, we further investigated a possible cause. If no such cause was apparent, the full form was labelled a missed opportunity. In the final step, we checked the ratio of contractions to total words within NFSs by NS and NNS status. The results of the aforementioned analyses follow.

Results

Given the small sample size of the study (N1 = 30 for NNSs and N2 = 15 for NSs), the results should be considered suggestive rather than definitive. Nonetheless, some patterns can be detected.

Quantitative analyses

Table 3 depicts the ratio of total number of words contained in nativelike formulaic sequences to the total number of words within each transcript for both NSs and NNSs by the final score awarded to the live test. The direction of the one-way ANOVA (rating) with four levels (A, B, C, D) is as expected: the higher the score, the higher the ratio of NFS production to total number of words. However, as Table 4 depicts, test takers who earned a C were largely grouped in one place. Table 5 shows that the correlation between the ratio of production of NFSs in Task 3 alone and the final score obtained on the test is moderate. In other words, this one evaluation criterion on this one task does not strongly predict the final score of the test.

Table 3: Ratio by final score

Final rating	Mean	N	Standard deviation	Minimum	Maximum
A	0.450	4	0.1291	0.3	0.6
B	0.400	17	0.1173	0.2	0.6
C	0.371	7	0.1254	0.1	0.5
D	0.250	2	0.0707	0.2	0.3
Total	0.390	30	0.1213	0.1	0.6

Table 4: Plot of ratio of NFSs by final score, NNSs

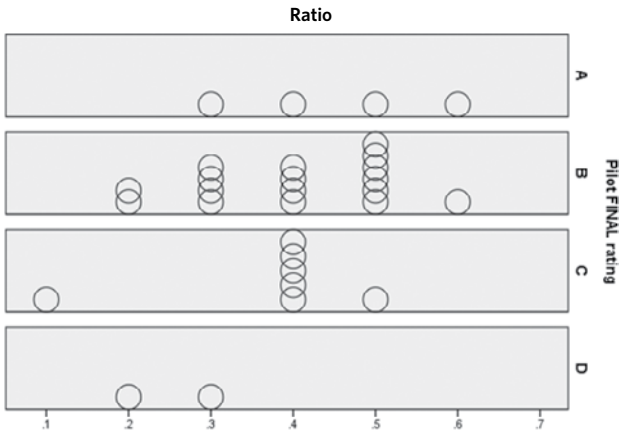


Table 5: Ratio by final score, NNSs only

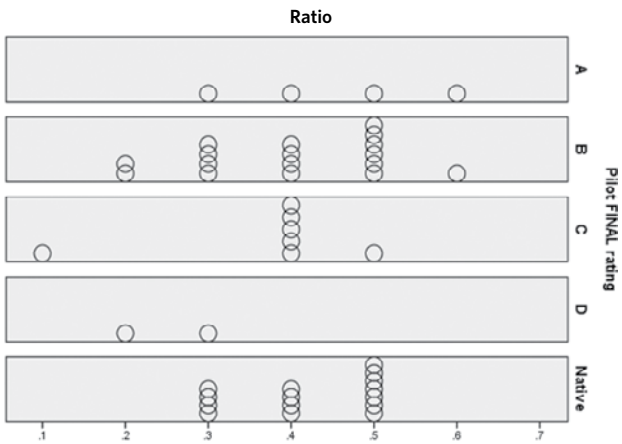
	ANOVA table				
	Sum of squares	**df**	**Mean square**	**F**	**Sig.**
Ratio between groups (Combined)	0.058	3	0.019	1.354	0.279
Final rating within groups	0.369	26	0.014		
Final rating total	0.427	29			

Analyses conducted to answer the second research question can be found in the following paragraphs. When NSs of English are included in the analysis as a comparison, we find that there is no statistically significant difference between their ratios of NFS production to total number of words and those of the NNSs (see Table 6). The NSs' means fall between the means of NNSs who received As and Bs on ITASA (see Table 7).

Table 6: Ratio by final score, NSs and NNSs

Final rating	**Mean**	**N**	**Standard deviation**	**Minimum**	**Maximum**
A	0.450	4	0.1291	0.3	0.6
B	0.400	17	0.1173	0.2	0.6
C	0.371	7	0.1254	0.1	0.5
D	0.250	2	0.0707	0.2	0.3
Native	0.420	15	0.0862	0.3	0.5
Total	0.400	45	0.1108	0.1	0.6

Table 7: Plot of ratio by final score, NSs and NNSs



The final analyses were run to detect any patterns in the use of contractions. The mean ratios are statistically significantly different (p-value = 0.008; t-statistic = −2.2776 with 43 degrees of freedom (df)). The NNSs have a statistically significantly lower mean ratio. The mean ratio for NSs was 26%; for NNSs the mean ratio was 15%. The NNSs were 11% lower, with a 95% confidence interval of (−19%, −3%). This means that NNSs significantly underused contractions in the base corpus compared to the reference corpus (Tables 8, 9, 10).

Table 8: Plot of contractions by speaker status

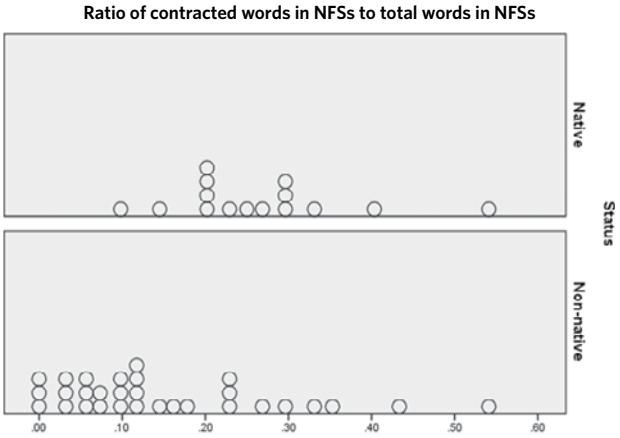


Table 9: Contractions to total words in NFSs by speaker status

Status	**Mean ratio**	**N**	**Standard deviation**
Native	0.2635	15	0.10727
Non-native	0.1527	30	0.13445
Total	0.1896	45	0.13553

Table 10: T-test for equality of means

t-test for equality of means of contractions to total word in NFSs by native and NNSs						
					95% confidence interval of the difference	
t	**df**	**Sig. (2-tailed)**	**Mean difference**	**Std. errror difference**	**Lower**	**Upper**
−2.776	43	0.008	−0.11081	0.03992	−0.19132	−0.03030

Summary of quantitative results

Statistical analyses suggest that the ratio of NFSs alone is not a strong predictor of the final ITASA score. Furthermore, the non-native population under study did not significantly differ in the ratio of NFS production from native-English-speaking TAs at the same institution. However, NNSs used significantly more carefully pronounced function words than did the NSs.

Qualitative analyses

Let us explore further the differences in the production of contractions between the native and non-native samples, with special focus on the way NNSs do and do not produce

contractions. Table 8 reveals that while approximately one third of the NNSs produced a comparable number of contractions to NSs; two thirds produced much less. Table 11 depicts examples of the occurrences of NFSs in COCA based on their pronunciations.

Table 11: COCA frequencies of contractions used and missed by NNSs

NFS used by test taker	COCA total # of occurrences	NFS not used by test taker	COCA total # of occurrences
You're welcome	2,431	You are welcome.	124
I can't do anything	222	I cannot do anything	11
I'm fine	415	I am fine	21
How are ya?	13	How are you?	5,321
How's that sound?	5	How does that sound?	21
We are not going to	775	We're not going to	3,837
		We're not gonna	190
		We are not gonna	5
		We aren't going to	158
We are going to	5,301	We're going to	40,089
		We're gonna	2,735
		We are gonna	55
I will try to	249	I'll try to	571

Frequently, when NNSs did use contractions, they used them in FSs that are very widely used. One example is 'You're welcome', which is both very common and nearly 20 times more frequent than its carefully pronounced counterpart. Other times, as in the case of 'I can't do anything' or 'I'm fine', the FSs themselves may not be very common but the ratio of occurrence of the contracted and full forms is equally notable, 20 and 19 times more frequent, respectively.

The fourth and fifth lines in the table show FSs whose pronunciation the test takers reduced, despite the fact that the reduced versions are less common in COCA than the carefully pronounced versions.

The final three formulaic sequences presented in the table were pronounced with careful pronunciation even though the reduced, contracted version is at least twice as common as the full pronunciation. The first two categories we coded as successful attempts at reduction. The third category was coded as a missed opportunity. Although clearly no points were awarded for missed opportunities, no points were lost either.

Discussion

The answer to the first research question, as to the degree to which the volume of NFSs produced in Task 3 Office-Hour Role Play contributes to the final rating, is not very surprising. The fact that a single evaluation criterion such as the production of NFSs as observed in a 2–3-minute task within a 14–15-minute test is not highly predictive of the final score is to be expected. As we know from the last decade of linguistics research into the area of formulaic language, the boundaries between the lexical and the grammatical are not as neatly defined as was once thought (see Ellis 2012, Schmitt 2004, and others). Therefore, perhaps the fact that evaluators in an earlier study about this test (Plough et al 2010) made frequent comments about the test takers' production of grammar and vocabulary, and that this study showed that a lexico-grammatical category is moderately correlated with the final score, are not incongruous findings. The findings seem to suggest that test takers who receive As, Bs and Cs tend to produce a volume of NFSs that is comparable to NSs' production of such sequences. In order to confidently distinguish the passing test takers from the ones who are not approved for ITA duties, the other evaluation criteria on the rating scale are needed.

The second research question asks whether the Task 3 Office-Hour Role Play elicits NFSs similarly to the real-life setting. In terms of ratio of words produced within a sequence to total number of words produced, the answer is yes. This suggests that the task meets an important criterion of performance tests in that its design not only simulates the real-life setting in which TAs may work in the future, but also elicits language that is similar to that produced by native-speaking TAs in a similar situation. This seems to be the case for the production of NFSs at least. The actual expressions used by NNSs and NSs are frequently identical in their interactional nature ('gonna', 'have a good day/weekend/time', 'come in', 'and stuff', 'how about you', 'how are you doing', 'have a seat', 'have any difficulties/concerns', 'is that helpful', etc.) and in their transactional nature ('a lot of information', 'if you have any questions/any other questions', 'because of', 'first of all', 'focus on', 'go through', 'I don't know/dunno', 'at least', 'based on', 'I mean', 'I think', 'I want you to', 'in fact', 'look at', 'make sure', etc.). TAs, regardless of their first language, perform many of the same speech acts using the same language, as they interact with students.

This study does, however, illuminate one difference in the pronunciation of many of these formulaic expressions. NSs of English used statistically significantly more contractions per words produced than did NNSs. So what can be the reason for this phenomenon? Perhaps it is that the speakers' interactional agenda differed. While the NSs were aware of being recorded, they were simply performing their teaching duties like many times before. The NNSs, on the other hand, were merely pretending to do the same. Although the testing situation in this study had no stakes, test takers may have inadvertently felt that they needed to monitor their speech. They may have wanted to sound especially comprehensible for the evaluators and thought that the best way to accomplish this was to use clear rather than reduced pronunciation in many cases. Another likely reason for this phenomenon is that learners prefer the full forms of verbs (Bardovi-Harlig 2012) even when not operating within the constraints of a test. They seem to have memorised conventional expressions as a whole, such as 'You're welcome' and 'I'm fine'. However, classroom-taught adult NNSs often do not store strings of words as a whole the way NSs of a language or children learning a second language do (Wray 2002). Instead, adults parse the formulaic sequences they are exposed to and store them as separate words. They only reassemble them at the time of language production. Therefore, if some of these sequences only become combined at the time of speaking, they are less likely to be reduced.

Limitations

The most notable limitation of the present study is the small sample size – 15 NSs and even 30 NNSs are too few to draw broad conclusions. Therefore, the results have to be interpreted with caution. The study should be repeated with a larger sample size. A further observation about the non-native sample is that it is very able. In a sense it can be considered truncated because only a subset of the non-native speaking population is permitted to take this exam: those who gain admittance to the university through meeting minimum English-language requirements on a placement test such as the Michigan English Language Assessment Battery (MELAB®). The claim that this is an able sample is supported by the fact that only 7% of the participants received the lowest score, D, and another 23% received a C.

Another limitation of this study concerns the fact that we attempted to establish a link between, on the one hand, a holistically scored test with multiple evaluation criteria and, on the other hand, one evaluation criterion about one task within a test. Statistical analyses have tried to correct for this fact; nevertheless, any inferences have to be drawn carefully on such an indirect link. Furthermore, NFSs were extracted from a transcript rather than video or audio. While reading a transcript may illuminate certain aspects of language, such as grammatical accuracy, it may also obscure other, more performative elements, such as pauses and thought groupings.

Acknowledgements

We would like to thank statistician Barry DeCicco for his significant contribution to the statistical analyses contained in this paper. We would also like to thank the following individuals whose comments, suggestions, and insights have strengthened our research: Jayanti Banerjee, Damir Cavar, Malgorzata Cavar, Natalie Nordby Chen, Nick Ellis, Ken Guire, Nabila Khan, Fabiana MacMillan, Brian Porter-Szucs, Ute Romer, Norbert Schmitt, Wendy Summers and Daniel Walter.

References

Anthony, L (2011) *AntConc (Version 3.2.2)*, Tokyo: Waseda University, available online: www.antlab.sci.waseda.ac.jp/

Bailey, K M (1983) Foreign teaching assistants at US universities: Problems in interaction and communication, *TESOL Quarterly* 17, 308–310.

Bardovi-Harlig, K (2012) Pragmatic routines, in Chapelle, C A (Ed) *The Encyclopedia of Applied Linguistics*, Oxford: Wiley-Blackwell.

Briggs, S L and Hofer, B (1991) Undergraduate perceptions of ITA effectiveness, in Nyquist, J D, Abbott, R D, Wulff, D H and Sprague, J (Eds) *Preparing the Professoriate of Tomorrow to Teach: Selected Readings in TA Training*, Dubuque: Kendall-Hunt Publishing Company, 435–445.

Briggs, S L, Madden, C G and Myers, C L (1994) Using performance assessment methods to screen ITAs, in Madden, C G and Myers, C L (Eds) *Discourse and Performance of International Teaching Assistants*, Alexandria: TESOL, 63–80.

Brown, G and Yule, G (1983) *Discourse Analysis*, Cambridge: Cambridge University Press.

Davies, M (2008) *The Corpus of Contemporary American English: 450 Million Words, 1990–present*, available online: corpus.byu.edu/coca/

Ellis, N (2012) Formulaic language and second language acquisition, *Annual Review of Applied Linguistics: Topics in Formulaic Language* 32, 17–44.

Nattinger, J R and DeCarrico, J S (1992) *Lexical Phrases and Language Teaching*, Oxford: Oxford University Press.

Plough, I C and Bogart, P H S (2008) Perceptions of examiner behavior modulate power relations in oral performance testing, *Language Assessment Quarterly* 5 (3), 195–217.

Plough, I C, Briggs, S L and Van Bonn, S (2010) A multi-method analysis of evaluation criteria used to assess speaking proficiency, *Language Testing* 27 (2), 235–260.

Schmitt, N (2004) *Formulaic Sequences: Acquisition, Processing and Use*, Amsterdam/Philadelphia: John Benjamins Publishing Company.

Simpson, R C, Briggs, S L, Ovens, J and Swales, J M (2002) *The Michigan Corpus of Academic Spoken English*, Ann Arbor: The Regents of the University of Michigan.

Wiggins, G (1989) A true test: Toward more authentic and equitable assessment, *Phi Delta Kappan* 70 (9), 703–713.

Wray, A (2002) *Formulaic Language and the Lexicon*, Cambridge: Cambridge University Press.

Dimensionality and factor structure of an English placement test

DANIEL WALTER RESEARCH GROUP, CAMBRIDGE MICHIGAN LANGUAGE ASSESSMENTS, USA
JASMINE HENTSCHEL UNIVERSITY OF MICHIGAN, USA

Background

The Cambridge Michigan Language Assessments (CaMLA) English Placement Test (EPT™) is designed to quickly and reliably group English as a Second Language (ESL) students into homogeneous ability levels in a wide variety of contexts. It is used by universities, colleges, language programmes, and businesses to evaluate students' and employees' ability to use English for language programme placement purposes or as part of hiring processes.

Originally introduced in 1972, the EPT is a measure of general receptive language proficiency aimed at language learners whose English proficiency ranges from high beginner to low-advanced, or Levels A2 to C1 on the Common

European Framework of Reference (CEFR) (Council of Europe 2001). In 2012, CaMLA revised the EPT to renew the content of the three original forms of the test (A, B, and C) while keeping the test format and score interpretation consistent so that users of older versions could move seamlessly to the new test forms. The result of this revision process was the construction of three new unique forms (D, E, and F) that contain sections parallel in difficulty across forms with items targeted at a range of ability levels. The forms were also designed to contain an even balance of domains, topics, and item sub-skills. To facilitate the use of form D, E, and F test scores for previous EPT users, a concordance table has been prepared to link forms A, B, and C and the new forms (available online: cambridgemichigan.org).

After the conclusion of the revision project, CaMLA began collecting live test taker data from the new forms for work on test validation research. As a change from previous views on validity, rather than viewing test validation as an argument towards an absolute standard of validity, current theory views it as context-dependent based on test construct (Weir 2005). Language testing experts present test validation as the cumulative result of multiple claim-based validity arguments about a particular test (Chapelle, Enright and Jamieson 2008, Kane 2006:22–23). For a placement test like the EPT, validity arguments often include arguments about score reliability, generalisability, interpretability, and accuracy of candidate placement. This paper is one of several such claim-based validity arguments that will be made for the EPT.

Rather than giving separate sub-scores for each section of the test, the EPT is designed to yield a single composite score that can be used as a measure of general receptive English language proficiency. The present study employs Confirmatory Factor Analysis (CFA) on test taker data from form F using four possible models to provide justification for reporting scores using a single scale instead of multiple subscales based on different language skills.

Literature review

Linguists have debated the componentiality of language ability for decades. The unitary model of linguistic proficiency took prominence for some time, particularly with the adoption of item response theory frameworks, which often require unidimensionality for test items being analysed. Papers such as Oller's 1983 analysis of the University of California, Los Angeles (UCLA) English as a Second Language Placement Examination (ESLPE) argued for the unitary factor model. In this study, he analysed the test as administered to four separate testing populations and concluded that a single English proficiency factor underpinned the ESLPE. In that same anthology, however, Oller discussed the two extremes of dimensionality interpretation, with a completely unitary view of proficiency on one side and 'an uncountable number of unrelated distinct elements which disallow any sort of global evaluation' (Oller (Ed) 1983:xii) on the other, concluding that 'no one has ever really believed either of these two options due to their extremity'. His implication was that a model could be found that harmonised both the unity and diversity of language skills.

The communicative language ability framework created by Bachman (1990:81) provided a theoretically based framework to explain multidimensionality in language proficiency, moving away from the abstract term *proficiency* towards a more pragmatic 'ability to use language communicatively', proposing that a wide variety of linguistic, psychological, and physiological characteristics make up communicative language ability. The framework's comprehensiveness, one of its primary strengths, also causes difficulty in implementation in a test construction setting (O'Sullivan (Ed) 2011:26). Chalhoub-Deville (1997:8) argued that:

> Assessments that are typically constructed in a given context will not include all the features depicted in [communicative language ability]; only those aspects highlighted by the variables operating in that context will be salient.

This view of language ability as being context dependent emerged in tandem with claim-based validity arguments, which led to a shift in focus for research. Many researchers began to focus less on investigating the nature of proficiency as a theoretical framework and more on validating test constructs by investigating the particular factor structure of the tests being examined. In 1998, Bae and Bachman (1998:384) wrote that 'applied linguists currently believe that language competence consists of multi-level components (such as grammar, pronunciation, organisation) and four skills (listening, reading, speaking, and writing)', and used standard error of measurement (SEM) to investigate the factor structure of the Korean Listening Test and Korean Reading Test, with individual tasks grouped together for the analysis. Their conclusion was that the best fit for the data was a two-factor solution of listening and reading as separate but correlated latent variables.

Over the next 15 years, a number of investigations of test factor structure concluded that correlated-trait models best explained the variance found in test items. Shin's (2005) SEM analysis of the Test of English as a Foreign Language (TOEFL®) found the best-fitting model was a second-order factor model with three lower-order factors: listening, writing, and speaking. A 2009 Internet-based Test of English as a Foreign Language (TOEFL iBT®) study by Sawaki, Stricker, and Oranje aimed to validate the TOEFL's score reporting procedures by examining factor structure using item-level factor analysis on a polychoric correlation matrix. This allowed it to offer highly granular insights into the behaviour of individual items. The best fit found in that study was a four-factor second-order model with reading, listening, speaking, and writing trait factors under a general proficiency factor. In'Nami and Koizumi's (2012) study of the Test of English for International Communication (TOEIC®) factor structure found a two-factor correlated model best fitted the TOEIC's construct, with reading and listening as the correlated latent factors.

EPT construct

By 2010, Bachman and Palmer (2010:56) 'conceptualise[d] "language skills" as the contextualised realisations of the capacity for language use in the performance of specific language use tasks'. To that end, the EPT construct narrows its proficiency framework from the ability to perform any language task in any circumstance to a much more narrowly defined *receptive language proficiency*. This is operationally

defined as the ability to successfully answer speeded multiple-choice questions that test listening comprehension, knowledge of grammatical formulation, vocabulary range and depth, and reading comprehension. Each item type targets different language interactions and domains, enabling test takers to demonstrate skills in a variety of contexts.

The test contains 80 questions and is administered in 60 minutes. The two listening item types are played via a digital audio recording, which is intended to ensure comparability between test administrations. Table 1 details the different item types and indicates the number of each on one EPT form.

Method

As part of the revision process, CaMLA conducted pilot testing before compilation of the new EPT forms. Statistics from pilot testing were used to identify unsatisfactory items and ensure that new forms would be at approximately the same level of difficulty. To determine whether all items shared similar factor structure, explanatory factor analysis was performed on the four pilot forms. Using the open-source statistical software package R, a tetrachoric correlation matrix was calculated, and an oblique rotation was applied to that matrix. Although the results of those analyses are probably less stable due to the small N sizes of 109, 95, 205, and 71, they revealed that a one-factor or a two-factor model could adequately explain most of the variation of the data. Finding similar factor structures led us to conclude that all three new forms would share the same structure.

Because forms D, E, and F were so recently released for use, the amount of data needed to perform meaningful analyses was only collected for form F at this time. However, because the different forms were carefully constructed to be parallel in both content and psychometric criteria, the results can be generalised across all three. Form F data was gathered from administrations given to 314 test takers over the course of six weeks. The information for two test takers was removed from the dataset: one because the test taker answered no questions, and the other because the test taker's answers were too faint to be accurately transcribed. This left a total sample of 312 test takers.

Table 1: EPT item types

Item type	Description	Number of items
Listening questions	Listen to a short question and select the best response from three answer choices	10
Listening dialogues	Listen to a conversation between two speakers and answer a question about the exchange by choosing from three answer choices	15
Grammar	Read a short dialogic exchange between two speakers in which part of a turn has been omitted and select which of four answer choices best completes the exchange	20
Vocabulary	Read a single sentence from which one word has been omitted and select which of four answer choices best completes the sentence	20
Sentence level reading	Read a single sentence and then answer a comprehension question by selecting one of four answer choices	5
Reading passage	Read a passage and answer comprehension questions about it by selecting from four answer choices	10 (with 2 passages)

Test takers represented a wide variety of backgrounds and English language education. Sixty-five percent were female and 34% male, with 1% failing to answer. The test takers ranged in age from 15 to 60, following the distribution shown in Figure 1. They were also asked to report the length of their English language study, the summary of which is presented in Figure 2. Test takers ranged from a reported lack of formal English instruction to having studied English for over 20 years, which implies a diverse level of experience with the English language.

In an effort to provide external evidence of proficiency, test takers were asked to provide results from previous standardised English language examinations they had taken. However, too few test takers responded for this information to be usefully analysed.

The dataset included test takers from 22 countries. Table 2 shows the distribution of test takers from the most common countries of origin, with all other countries encompassed by the 'Other 13 countries' category.

Figure 1: Age of test takers in years

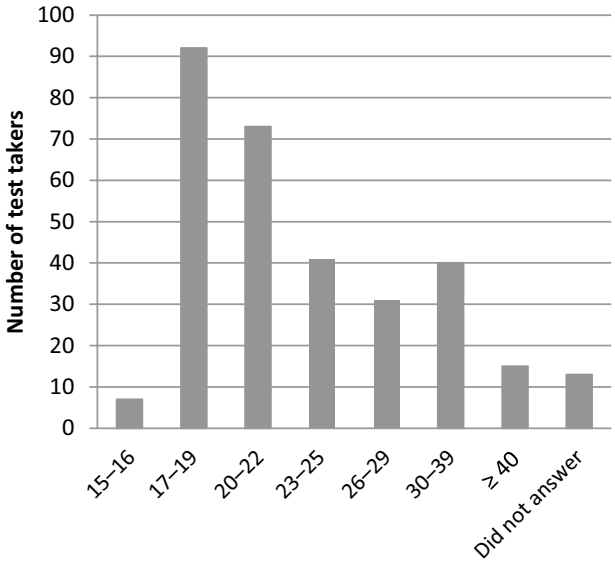


Figure 2: Length of test takers' English study

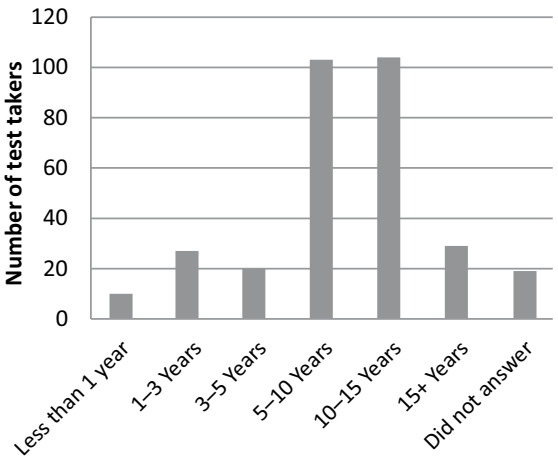


Table 2: Test taker countries of origin

Country of origin	% of sample
Italy	20.51
Brazil	15.38
Japan	12.18
South Korea	10.58
France	8.65
Spain	8.33
Russia	6.09
Taiwan	4.49
Germany	4.17
Other 13 countries	9.62

The diversity of the form F test taking sample across gender, age, linguistic background, and length of English study is very similar to that of data collected during pilot testing, conducted at 13 different test centres across the USA and Canada. Most centres were already EPT users or planned to become users; this indicates that the results of this study should be generalisable to other EPT test taking populations.

Normality assumptions

The CFA and data analyses were performed using LISREL 8.8 (Jöreskog and Sörbom 2007). Because Maximum Likelihood parameter estimation assumes multivariate normality of the data, univariate and multivariate normality assumptions were checked using LISREL's PRELIS capabilities. Descriptive statistics for the six item types of the EPT are listed in Table 3.

Table 3: Descriptive statistics

Item types	Mean	St. dev.	Skewness	Kurtosis	Min.	Max.
Listening question	7.01	2.42	−0.53	−0.74	0	10
Listening dialogue	11.13	3.34	−0.81	−0.21	1	15
Grammar	14.18	4.24	−0.73	−0.32	4	20
Vocabulary	13.58	4.37	−0.57	−0.32	0	20
Sentence level reading	3.46	1.56	−0.86	−0.36	0	5
Reading passage	6.10	3.23	−0.53	−1.04	0	10

It is impossible to achieve multivariate normality without univariate normality (Thompson 2004:122), so univariate measures were inspected first. The EPT subscores for each item type show some skewness and kurtosis, with the sentence level reading displaying statistically significant skewness and kurtosis values, and the reading passage section displaying a significant kurtosis value (see Table 4). The negative skewness for every item type implies that the entire test was negatively skewed, potentially indicating that the test was slightly easy for the population. This is partially corroborated by the fact that the mean for each item type is at least half of the maximum possible number of points. In addition, both of the reading item types have X^2 values that indicate significant non-normality.

In addition to univariate measures, Prelis multivariate skewness and kurtosis tests revealed a chi-square of 93.20 with a p-value of 0.00, indicating statistically significant deviation from multivariate normality for the entire dataset.

There are two primary methods for performing structural equation modelling (SEM) calculations on non-normal data. One common method is to use an estimation that has no normality assumptions. In particular, weighted least squares (WLS) estimation methods are often used. However, Olsson, Foss, Troy, and Howell (2000:557) found that WLS algorithms are even unstable with sample sizes of at least 1,000, which has been confirmed elsewhere (Ullman 2006:43). Since the sample used for this analysis is 312, using WLS algorithms would introduce additional complications for model identification.

Another method to perform SEM on non-normal data uses some linear transformation of the data that preserves score order while bringing it close to randomness (Kline 2011:177). Possible transformations 'include square root, reciprocal, logit, or probit' transformations (Schumaker and Lomax 2004:33). Consequently, the data were transformed with LISREL's PRELIS normal scores using the LISREL method (Jöreskog, Sörbom, du Toit and du Toit 1999) for the six observed variables. As Table 5 indicates, the transformed data had a multivariate skewness and kurtosis chi-square of 2.46 and a p-value of 0.29, which implies the transformed data was sufficiently multivariate normal for SEM analysis.

Table 4: Skewness and kurtosis of sub-skills

Item types	Skewness Z-score	p-value	Kurt. Z-Score	p-value	χ^2	p-value
Listening question	−1.28	0.20	−1.91	0.06	5.29	0.07
Listening dialogue	−1.17	0.24	−1.72	0.09	4.33	0.12
Grammar	−0.28	0.77	−1.01	0.31	1.10	0.58
Vocabulary	−0.39	0.70	−0.62	0.53	0.54	0.76
Sentence level reading	−2.37	0.02	−5.27	0.00	33.33	0.00
Reading passage	−0.57	0.57	−3.41	0.00	11.97	0.00

Table 5: Skewness and kurtosis values for normalised data

	Skewness Z-Score	p-value	Kurt. Z-Score	p-value	χ^2	p-value
Normal scores	0.83	0.40	−1.33	0.18	2.46	0.29

Confirmatory factor analysis models

Four major model types were tested based on the different language skills the EPT assesses.

Single-factor model Since the EPT is described as a measure of general language proficiency and scores are considered part of a consistent test of proficiency, a model wherein each item type is part of a single factor was tested (see Figure 3).

Uncorrelated two-factor model In this model, the listening question and listening dialogue items load on a general listening factor, and the grammar, vocabulary, sentence level reading, and reading passage items load on a general reading factor. The reading and listening factors are uncorrelated (see Figure 4).

Correlated two-factor model Three models fall in this family due to the nature of the grammar items. Because they are written in the form of short dialogues, in some ways they behave more like listening than reading items. Three variations were proposed to correctly determine which is most accurate: a model with grammar as listening, a model with grammar as reading, or a model with grammar as both. Figure 5 compares the three models.

Because of the nature of SEM models, if the second-order factor has the covariance set to 1, there is no statistical difference between a correlated two-factor model and a three-factor second-order model (see Figure 6) with one general language proficiency factor that impacts listening and reading factors (see the third diagram in Figure 5).

Four-factor correlated model In this model, the four language skills tested on the EPT – listening, grammar, vocabulary, and reading – are posited as the underlying structures beneath the observed variables.

Results

Analyses were conducted in LISREL using SIMPLIS syntax on the raw data, the covariance matrix of which can be found in the Appendix. As the analyses were done, the four-factor model failed to converge, probably because the number of observations (6)(6+1)/2 provided too few degrees of freedom

Figure 3: Path diagram of single-factor model

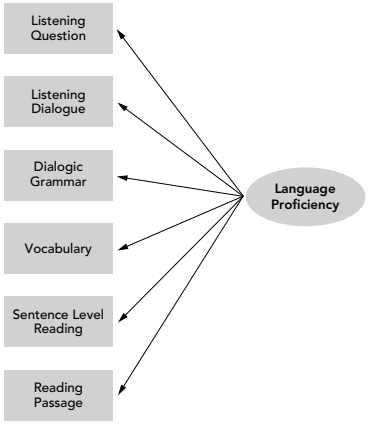


Figure 4: Path diagram of uncorrelated two-factor model

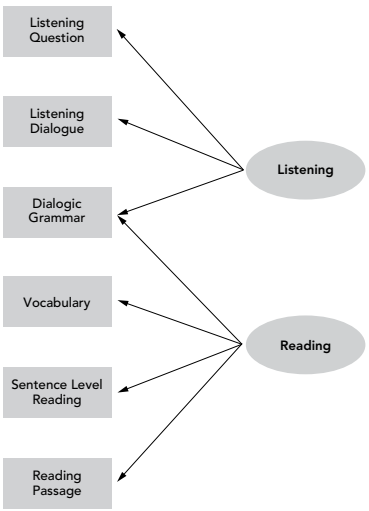
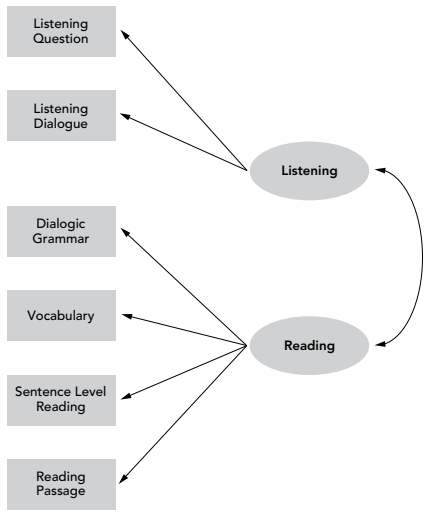


Figure 5: Path diagrams of three different correlated two-factor models



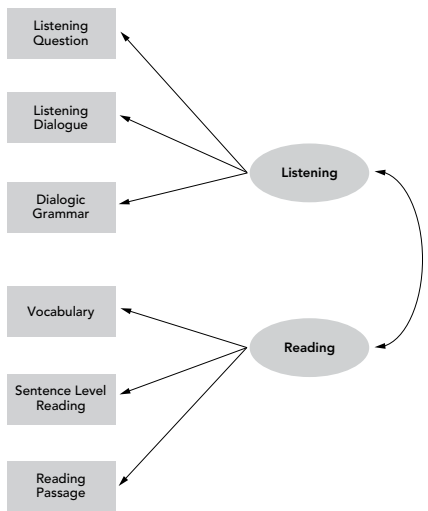
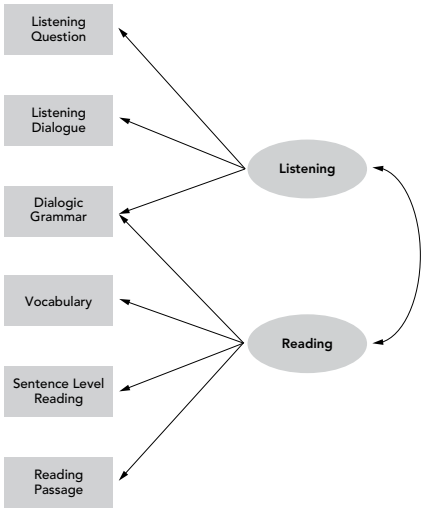


Figure 5 (continued)



for the model to be correctly identified. No other convergence failures or Heywood cases (negative estimated variance) presented themselves.

Fit indices

Fit indices for all models that did not fail to converge are listed in Table 6, including three variations of the two-factor correlated model: one with grammar as listening (GLIS), one with grammar as reading (GRDG), and one with grammar as both (GBOTH). Descriptions of the different fit indices can be found below.

df, X^2 The model X^2 value describes a central X^2 with the specified number of degrees of freedom (df). For the single factor model, the X^2 is 71.20 and the degrees of freedom are 9. These numbers can be used to discern how well the data fits the model, with a better fitting model producing a lower X^2 value (Kline 2011:199). According to this measure, the two-factor GBOTH model is the best fit.

GFI The Goodness of Fit Index 'estimates how much better the researcher's model fits compared with no model at all' (Kline 2011:207). Model fit usually ranges between 0 and 1; a higher number indicates better fit. According to this measure as well, the two-factor GBOTH model is the best fit.

NFI The Normed Fit Index evaluates model fit from 0 to 1, with a value higher than 0.95 indicating close fit (Schumaker and Lomax 2004:82). This index also identifies the two-factor GBOTH model as best fit.

Figure 6: Path diagram of second-order model

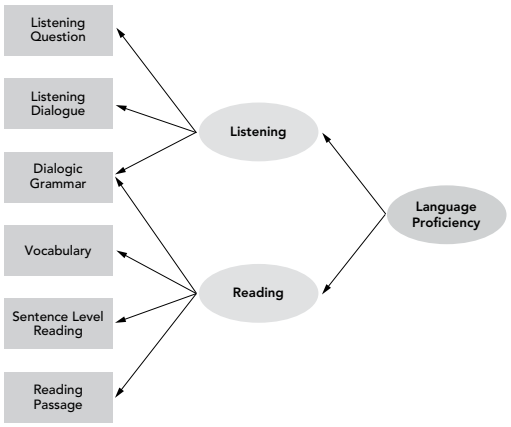
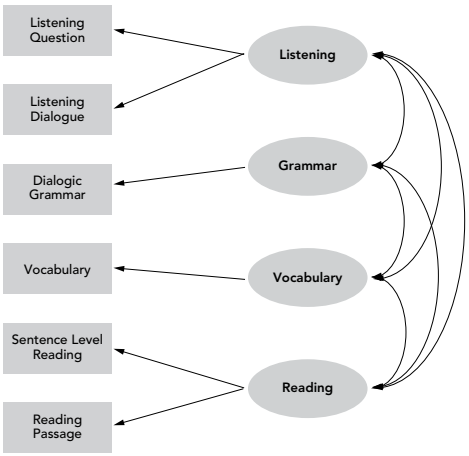


Figure 7: Path diagram of correlated four-factor model



AIC The Akaike Information Criterion is a method used 'to select among competing non-hierarchical models estimated with the same data' (Kline 2011:220). Lower values of the AIC are better. The two-factor GBOTH model is the best fit using this measure.

RMSEA The Root Mean Square Error of Approximation is a measure of badness-of-fit. Values of 0.10 or below are indicators of good model fit, and values below 0.05 of very good fit. If the larger value of the 90% Confidence Interval (CI) for the RMSEA is above the threshold being considered, then as with other hypothesis tests, 'the model warrants less confidence' (Kline 2011:206). Again, the best-fitting model is the two-factor GBOTH model.

The various fit indices are aligned for this dataset; the following models are recommended in increasing order

Table 6: Fit indices for successful models

Model	df	X^2	GFI	NFI	AIC	RMSEA	RMSEA 90% CI
Single-factor	9	71.20	0.92	0.96	103.51	0.16	0.13–0.19
Two-factor uncorrelated	8	257.34	0.84	0.87	203.14	0.26	0.23–0.29
Two-factor GLIS	8	27.34	0.97	0.99	52.07	0.09	0.05–0.12
Two-factor GRDG	8	49.34	0.95	0.97	77.99	0.13	0.10–0.17
Two-factor GBOTH	7	22.42	0.98	0.99	49.82	0.08	0.04–0.12

of model fit: two-factor uncorrelated model, single-factor model, two-factor correlated model with grammar as listening, two-factor correlated model with grammar as reading, and two-factor correlated model with grammar as both listening and reading.

The single-factor model obviously describes some of the variation found in the dataset, but it does not explain enough of the data to recommend adopting it. The model that best fits the data according to every index is the two-factor correlated model with both listening and reading affecting the grammar items. The GFI of 0.98 and NFI of 0.99 both indicate very good model fit. The RMSEA of 0.08 indicates good fit despite failing to reach the best-fit criterion of 0.05.

Conclusion

The best-fitting model has grammar loading on both the listening and reading factors, but on listening more than reading. This makes it difficult to split EPT scores into two distinct subscores of listening and reading, since the dialogic grammar items wouldn't clearly fit into one subsection or the other. Because a second-order model with three factors and a correlated two-factor model are statistically identical, the correlated two-factor model with grammar as both listening and reading could also be interpreted as a second-order factor model with a single factor (general receptive language proficiency) underlying two factors (listening and reading) that explain the variation in the data. Therefore, the EPT scale represents a measure of general receptive language proficiency supported by listening and reading sub-skills. This aligns with the reporting of EPT results as a composite score on a single scale.

Because the best-fitting model still did not reach a RMSEA of 0.05, it is possible that a different (in all probability, more complicated) model better fits the data. A further study with a much larger sample size that utilises a full item information factor analysis like Sawaki et al's TOEFL study (2009) would be able to investigate bifactor models and other more complex models of the EPT, which may better explain the factor structure of the test.

References

Bachman, L (1990) *Fundamental Considerations in Language Testing*, Oxford: Oxford University Press.

Bachman, L and Palmer, A (2010) *Language Assessment in Practice*, Oxford: Oxford University Press.

Bae, J and Bachman, L (1998) A latent variable approach to listening and reading: testing factorial invariance across two groups of children in the Korean/English Two-Way Immersion program, *Language Testing* 15 (3), 380–414.

Chalhoub-Deville, M (1997) Theoretical Models, assessment frameworks and test construction, *Language Testing* 14 (3), 3–22.

Chapelle, C, Enright, M and Jamieson, J (2008) Test Score Interpretation and Use, in Chapelle, C, Enright, M and Jamieson, J (Eds) *Building a Validity Argument for the Test of English as a Foreign Language*, New York: Routledge, 1–26.

Council of Europe (2001) *The Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

In'Nami, Y and Koizumi, R (2012) Factor structure of the revised TOEIC® test: A multiple-sample analysis, *Language Testing* 29 (1), 131–152.

Jöreskog, K G and Sörbom, D (2007) *LISREL 8.8.*, Lincolnwood: Scientific Software International, Inc.

Jöreskog, K G, Sörbom, D, du Toit S H C and du Toit, M (1999) *LISREL 8: New Statistical Features*, Lincolnwood: Scientific Software International Inc.

Kane, M (2006) Validation, in Brennan, R (Ed) *Educational Measurement*, 4th edition, Westport: Praeger Publishing, 17–64.

Kline, R (2011) *Principles and Practice of Structural Equation Modeling*, 3rd edition, New York: Guilford Press.

Oller, J (Ed) (1983) *Issues in Language Testing Research*, Rowley: Newbury House.

Olsson, U, Foss, T, Troy, S and Howell, R (2000) The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality, *Structural Equation Modeling* 7 (4), 557–595.

O'Sullivan, B (Ed) (2011) *Language Testing*, Houndmills: Palgrave Macmillan.

Sawaki, Y, Stricker, L and Oranje, A (2009) Factor structure of the TOEFL Internet-based test, *Language Testing* 26 (1), 5–30.

Schumaker, R and Lomax, R (2004) *A Beginner's Guide to Structural Equation Modeling*, 3rd edition, Mahwah: Lawrence Erlbaum.

Shin, S (2005) Did they take the same test? Examinee Language Proficiency and the structure of language tests, *Language Testing* 22 (1), 31–57.

Thompson, B (2004) *Exploratory and Confirmatory Factor Analysis*, Washington, DC: American Psychological Association.

Ullman, J (2006) Structural equation modeling: Reviewing the basics and moving forward, *Journal of Personality Assessment* 87 (1), 35–50.

Weir, C J (2005) *Language Testing and Validation*, Basingstoke: Palgrave Macmillan.

Appendix: Covariance matrix used in CFA calculations

	Listening question	Listening dialogue	Grammar	Vocabulary	Sentence level reading	Reading passage
Listening question	5.86					
Listening dialogue	5.94	11.17				
Grammar	7.39	9.85	17.98			
Vocabulary	6.39	9.7	12.74	19.09		
Sentence level reading	2.03	3.15	4.05	4.85	2.43	
Reading passage	4.5	6.85	8.22	9.03	3.35	10.44

698560612 9

To subscribe to *Research Notes* and download previous issues, please visit:
www.cambridgeenglish.org/research-notes

Contents:

For further information visit the website:
www.cambridgeenglish.org

Cambridge English
Language Assessment
1 Hills Road
Cambridge
CB1 2EU
United Kingdom

www.cambridgeenglish.org/help

ALTE Association of Language Testers in Europe

A DIVISION OF CAMBRIDGE ASSESSMENT