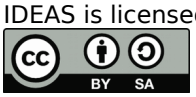
IDEAS

Journal of Language Teaching and Learning, Linguistics and Literature

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

Volume 6, Number 2, December 2018
pp. 49 – 61



Issued by English study program of IAIN Palopo

Validity, Reliability and Practicality Of The First Certification in English (FCE) and The Business Language Testing Service (Bulats)

Muhammad Ahkam Arifin
ahkam.arifin@parahikma.ac.id; ahkam.arifin@gmail.com
Institut Parahikma Indonesia

Received: 10 Oktober 2018; Accepted : 28 November 2018
URL : http://ejournal.iainpalopo.ac.id/index.php/ideas

Abstract

This paper begins with the test specifications of the two tests – the First Certification in English (FCE) and the Business Language Testing Service (BULATS). It will then go on to the evaluation of the test usefulness: reliability, (construct) validity, backwash, and practicality (Bachman & Palmer, 1996; see Kunnan, 2004 for a slightly different perspective). This paper explores the test specifications at the outset in that a test would be evaluated (as estimated) based on the test purpose and construct that it is trying to measure (Luoma, 2004). To begin the evaluation, the test (score) reliability would be evaluated first, for a test would not be considered valid if it is not reliable (Brown, 1996; but see Moss, 1994 when a test could be valid without reliability). Throughout this paper, the term “test(ing)” will be used more or less synonymously with “assess(ment)” and “measure(ment)”, in that Bachman and Palmer point out that in the field of language testing these terms have been very broadly defined “as the process of collecting information” to make decisions (2010, p. 20). (See Bachman, 1990; Cohen & Swedlik, 2010; Douglas, 2010 for the distinctions, e.g., a test is a tool for assessment.)

Keywords: evaluation, validity, reliability, practicality, first certification in English, business language testing service

Introduction

Both the FCE and BULATS are produced by the University of Cambridge Local Examinations Syndicate (UCLES), both of which offer either a paper-based or computer-based exam, covering the four skills[1]. This paper will focus upon the FCE paper-based speaking test and BULATS computer-based online speaking test. The FCE is intended for learners who want to live, work or study in an English-speaking environment (UCLES, 2015). The FCE is targeted at level B2 (Upper-Intermediate) on the CEFR[2], while the BULATS at all levels (UCLES, 2011). The BULATS is intended for "people at work or students studying business courses" (UCLES, 2011, p. 2). It takes 14 minutes for the FCE and 15 minutes for the BULATS to complete.

The FCE is composed of four sections: (1) an interview between the interlocutor[3] and each of the two candidates, (2) a presentation from each candidate, (3) a peer-peer interaction between the candidates, and (4) a discussion between the examiner and candidates. The BULATS, however, involves five parts; (1) interview, (2) reading aloud, (3) presentation about a work-related topic, (4) presentation with graphics, and (5) a communication activity by responding to questions on a specific situation.

The construct definitions of the FCE include grammatical and lexical resource, discourse management, pronunciation and interactive communication (please see Appendix 1A, 1B & IV for detailed definitions); the construct definitions include what construct / abilities / skills, that a test is intended to measure (Hughes, 2003; see Chappele, 1998 for three types of construct definitions). The BULATS focuses upon the student task achievement, coherence / discourse management, language resource, pronunciation, and hesitation / extent (see Appendix II). For the "reading-aloud" part, the constructs are focused upon the student overall intelligibility (pronunciation), individual sounds, stress, including rhythm, and intonation (see Appendix III).

The FCE employs analytic rating scales, for they set a number of criteria each of which has descriptors at the different levels of the scale, whereas the BULATS employs holistic rating scales, that is, to report an overall impression of the student ability in one score (Fulcher, 2003). The FCE interlocutor uses the holistic rating scale (see Appendix 1C), whereas the assessor uses the analytic one (Galaczi, 2008).

Regarding the interpretations of the test results, both the FCE and BULATS can be categorized into criterion-referenced testing (CRT) in that the student scores are interpreted in relation to one or more standards, objectives and other criteria, e.g., what they can and cannot do (Hughes, 2003; see Brown, & Hudson, 2002 for the development of CRT in response to NRT[4]). However, Wiliam argues that "the requirement that a criterion is useful for distinguishing

---

[1] The four skills: reading, writing, listening and speaking.

[2] CEFR stands for Common European Framework of Reference.

[3] In Cambridge ESOL terms, the "interlocutor" is the examiner who participates in the test and provides a global mark based on a holistic scale; while the nonparticipating examiner, called the "assessor", awards four analytical marks.

[4] , Norm-referenced testing (NRT)

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

levels of performance means that we have to use norms, however implicitly" (1993, p. 341).

Considering that both assess the student level of language ability without respect to any particular program or curriculum, both could be considered proficiency tests (Brown, 1995). However, both could also be regarded as admission tests, for the FCE is used for "entry to undergraduate programmes" [5] and BULATS is "for admission to study business-related courses" [6].

Reliability Evaluation

This section will in turn focus upon the reliability of the FCE and BULATS; reliability is simply defined as the "consistency of scoring or measurement" (Bachman & Palmer, 1996, p. 19). That is, the more similar scores the students have for the same test by taking it in different settings or situations, the more reliable the test is said to be (Hughes, 2003). One common way to analyse the reliability of a test is through "correlation", that is, a statistical indicator for the strength of relationship between two (or more) sets of measures which are considered to be related (Luoma, 2004). This relationship is then known as the correlation coefficient (Davies et al., 1999). Theoretically, values for correlation coefficients range between .00 and 1.00. While values close to zero indicate no relationship, values close to 1 means a perfect positive correlation (Hughes, 2003), notwithstanding neither extreme never occurs in practice. Carr (2011) argues that a reliability of .80 is generally set as a minimum level for high-stakes testing, whereas Lado (1961) claims that the speaking test ranging from .70 to .79 would be reliable enough.

The FCE speaking test has a reliability of .84[7], which could be considered to be quite high (Lado, ibid). Hacket (2002) reports that the overall reliability alphas of the BULATS vary from .95 to .96, while each section ranges between .85 and .92. Using Item Response Theory (IRT) (e.g., the Rasch model) [8], Jones (2000) shows that the standard error measurement (SEM) of the overall BULATS test is .33, whereas the overall FCE test has a SEM of 2.78 and of 1.50 for the speaking test[9]. The SEM concerns with the reliability of individual scores rather than the reliability of tests, that is, the estimation of how close the individual actual (or true) score; the variability caused by other factors (e.g., motivation or tiredness) is called error (Hughes, 2003; McNamara, 1996). Hence, statistically the BULATS could be said to have higher reliability than the FCE.

Throughout this paper the term "reliability" is synonymous with "dependability"[10], or what Brown (1990) calls "decision consistency".

[5] http://www.cambridgeenglish.org/exams/first/
[6] http://www.bulats.org/why-bulats
[7] http://www.cambridgeenglish.org/research-and-validation/quality-and-accountability/
[8] IRT is a general measurement theory. It assumes that for an item with a given level of difficulty, the probability that a test taker will answer correctly depend on their level of ability. Rasch model is one type of IRT (Carr, 2011).
[9] http://www.cambridgeenglish.org/research-and-validation/quality-and-accountability/
[10] UCLES (2013) uses both these terms more or less interchangeably.

Dependability has been used for CRT and reliability for NRT (e.g, Brown & Hudson, 2002; see also Ennis, 1999 for the use of “consistency” over “reliability”). Orr (2002) reveals that the raters for the FCE speaking test are found to not heed the same aspects of criterion, which then will result in giving different scorers. The same issue may apply to any other types of test (McNamara, 1996), which means that it could also apply to the BULATS. (Further research is needed.) However, inasmuch as the UCLES provides rater training, the reliability then could be enhanced in the sense that the raters would arguably give a similar score albeit on different occasions, and the similar score would also be given by another rater; the former is called “intra-rater reliability”, or “internal consistency” (Luoma, 2004), while the latter “inter-rater reliability” (Hughes, 2003).

Moreover, Bonk and Ockey (2003) report that returning raters will tend to move toward better consistency, as they get more experience. The use of rating scales (or rubric), more varied pattern of interaction in the FCE (as well as two examiners) and BULATS may also reduce the subjectivity of the scores that will affect the reliability. (See Bachman, 1990 for the problematic distinction between the so-called subjective and objective tests.) To conclude, Weigle`s claim that a holistic scale have weaker reliability than an analytic scale may not apply to all contexts (Weigles, 2002), particularly the speaking tests for both the FCE and BULATS.

Validity Evaluation

Secondly, both tests would be evaluated whether they accurately assess what they are intended to assess, this evaluation is called validity (Davies, 1990; Brown, 2005). Nonetheless, Green (2014) considers this definition to be classic, and goes on claiming that this “definition is now seen to be too limited and somewhat misleading” (2014, p. 75). *Standards for Educational and Psychological Testing* defines validity as “the degree to which evidence and theory support the interpretation of test scores entailed by proposed tests” (American Educational research Association et al., 1999). Carr (2011) also warns that we should not speak of validating a test *per se*, but we should speak of validating specific uses of a test. Throughout this paper validity is used synonymously with construct validity, given that the broader field of educational and psychological measurement and testing uses construct validity as “the whole of validity theory” (Shepard, 1993, p. 418; see Thorndike & Hagen, 1986, Cronbach & Meehl, 1995 for the past use of construct validity under validity).

To evaluate the degree of the construct validity of the FCE and BULATS tests, one essential type of validity evidence (not type of validity) to explore is content validity, also called ‘‘definition validity’’ and ‘‘logical validity’’ (Newman et al., 2006). A measurement could be said to have content validity if its content represents the full range of constructs (or knowledge, skills, abilities) that it is intended to cover (Alderson et al., 1995). Using observation checklists which contain a set of functions (see Appendix IV), O'Sullivan et al. (2002) report that the contents of the FCE have constituted a representative sample of those functions (or construct). For the BULATS, in the handbook a list of possible functions that will be tested in the speaking test is provided.

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

Thus, I assume the BULATS have content evidence validity, considering also that the BULATS is constructed by a group of language testing experts, who will have to provide argument regarding the type of underlying construct for each of the test item (or prompt). (Further research is needed).

A similar type of evidence is face validity. As a stakeholder, I believe that both tests have face validity in that they directly ask the learners to speak (Green, 2014), albeit the BULATS could be considered to be semi-direct, for there is no face-to-face interaction with an interlocutor (Fulcher, 2003; but see Carr, 2011 who finds the distinction of direct and indirect tests problematic). Other types of evidence are concurrent-validity and predictive validity, both are under "criterion-related validity" (Weir, 2005). Using predictive validity evidence by examining university students, Al-Musawi and Al-Anshari (1999, p. 389) claim that "the multivariate of the GPA from the scores on the FCE is very accurate". However, they do not focus solely on the speaking section. Weir (2005) points out that concurrent validation is concerned with the comparison of test scores with another measure of performance, e.g., another well-established test taken at the same time, teacher ranking of students, or student self-assessment.

Regarding the form of computer test of the BULATS, Chambers and Ingham (2011) report that test takers and examiners show overall positive feedback on the use of computer for the speaking test, and it could be considered to be type of evidence validity called "response validity" (Alderson et al., 1995). Nevertheless, they also find that some may find it more stressful, for they will not have an interlocutor support particularly when they need the questions to be repeated.

Washback and Practicality Evaluation

Messick claims that "for a fully unified view of validity" (1989, p. 18), social values and consequences of a test should be taken into considerations, and these consequences have been called "impact", that is, the effects of a test on "individuals, policies or practices, within classroom, the school, the educational system or society as a whole" (Wall, 1997, p. 291). It then leads to the creation of Code of Ethics for the International Language Testing Association (see Davies, 2003). Seemingly the most commonly discussed aspect of impact is backwash[11] (or washback). Backwash has been generally defined as the beneficial or harmful effects the tests have on teaching and learning (e.g., Hughes, 2003; Alderson & Wall, 1993).

Comparing the BULATS and FCE, I believe that the latter has more beneficial backwash because it is more authentic than the former, which can be considered to be semi-direct (Carr, 2011). If I were teaching to prepare my students for the FCE, I would create more opportunities for my student to have more discussion among them in my classroom, for the FCE provides peer-peer interaction and discussion. However, the FCE also may bring harmful impact to the extent that it may create subjectivity as the examiners can directly identify the students (lack of privacy and confidentiality), which then may lead to the fairness issue; privacy and confidentiality are basic rights

[11] Although washback is rarely found in dictionaries, it has been commonly used in applied linguistics today.

to the test takers (Bachman & Palmer, 1996).

Regarding practicality, in classroom contexts the FCE could be considered to be more practical in that teachers may need to provide computers and the internet connection for teaching the BULATS. Nonetheless in administrating the test, the BULATS could be more practical as it will not require the examiners to be in the testing room, rather it uses computer to record the student voice. At any rate, Bachman and Palmer (1996) cautions that a test should not consider practicality to be less important than other qualities, instead all qualities should be weighed the same.

References

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(1), 115–129.

Alderson, J.C., Clapham, C & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

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

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

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford. University Press.

Bonk, W. J. & G. J. Ockey. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing 20*(1), 89–110.

Brown, J. D. (1990). Short-cut estimators of criterion-referenced test consistency. *Language Testing*, *7*(1), 77-97.

Brown, J. D. (1995). *The elements of language curriculum: A systematic approach to program design*. Boston: Heinle & Heinle

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York, NY: McGraw Hill College.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford: Oxford University Press.

Chambers, L and Ingham, K. (2011). The BULATS online speaking test. *Research Notes 43*(1), 21–25.

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In Bachman, L. and Cohen, A., (eds), *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press, 32–70.

Cohen, R. J., & Swerdlik, M. E. (2010). *Psychological testing and assessment: An introduction to tests and measurement* (7th ed.). New York: McGraw-Hill.

Cronbach, L. J., & Meehl, P. E. (2010). Construct validity in psychological tests. *Psychological Bulletin, 52*(1), 281-302

Davies, A. (1990). P*rinciples of language testing*. Oxford: Blackwell.

Davies, A. (2003). Three heresies of language testing research. *Language Testing 20*(4), 355-368.

Davies, A., Brown, A., Elder, C. and Hill, K. (1999). *Dictionary of Language Testing.* Cambridge: Cambridge University Press.

Douglas, D. (2010). *Understanding language testing*. New York: Routledge

Ennis, R. H. (1999). Test reliability: a practical exemplification of ordinary language philosophy. In R. Curren (ed.), *Philosophy of education*. Urbana, IL: The Philosophy of Education Society, 242–48.

Fulcher, G. (2003). *Testing second language speaking*. London, UK: Pearson-Longman

Galaczi, E. D. (2008). Peer–peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly*, *5*(2), 89-119.

Green, A. (2014). *Exploring language assessment and testing: language in action*. New York: Routledge.

Hackett, E. (2002). Revising the BULATS standard test. *Research Notes 8*(1), 7-10

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

Jones, N (2000). BULATS: a case study comparing computer-based and paper-and-pencil tests. *Research Notes 3*, 10–13.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic and C. Weir (Eds.), *European language testing in a global context*. Cambridge: Cambridge University Press.

Lado, R. (1961). *Language testing*. London: Longman.

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

McNamara, T. (1996). *Measuring Second Language Performance*. London: Longman.

McNamara, T. F. & Roever, C. (*2006). Language testing: The social dimension*. Oxford: Blackwell.

Messick, S. (1989). Validity. In R. Linn (ed.), Educational Measurement. New York: Macmillan, pp. 13–103

Moss, P. A. (1994). Can there be validity without reliability?. *Educational researcher*, *23*(2), 5-12.

Newman, I., Newman, C., Brown, R., & McNeely, S. (2006). *Conceptual statistics for beginners* (3rd ed.). Lanham, MD: University Press of America.

Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System*, *30*(2), 143-154.

O'Sullivan, B., Weir, C. J. and Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing 19*(1): 33–56.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*(1), 405–. 450.

Thorndike, R. L., & Hagen, E. P. (1977). *Measurement and evaluation in psychology and education*. New York: Wiley.

UCLES (University of Cambridge Local Examinations Syndicate). (2015). *Cambridge English First: Handbook for Teachers*. Cambridge: University of Cambridge Local Examinations Syndicate.

UCLES (University of Cambridge Local Examinations Syndicate). (2011). *BULATS Business Language Testing Service: Information for Candidates.* Cambridge: University of Cambridge Local Examinations Syndicate.

UCLES (University of Cambridge Local Examinations Syndicate). 2013. *Principles of good practice quality management and validation in language assessment: Validity, reliability, impact, practicality.* Cambridge: University of Cambridge Local Examinations Syndicate.

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

Wall, D. (1997). Impact and washback in language testing, in C. Clapham and D. Corson (eds.), *Encyclopedia of Language and Education*, 291–302

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Weir, C. J. (2005). *Language testing and validation: An evidence based approach*. Houndgrave, Hampshire: Palgrave MacMillan.

Wiliam, D. (1993). Validity, dependability and reliability in national curriculum assessment. *The Curriculum Journal*, *4*(3), 335-350.

Appendix 1A Analytic Rating Scales for FCE Speaking Test

Overall Speaking scales

	Grammatical Resource	Lexical Resource	Discourse Management	Pronunciation	Interactive Communication
	• Maintains control of a wide range of grammatical forms and uses them with flexibility.	• Uses a wide range of appropriate vocabulary with flexibility to give and exchange views on unfamiliar and abstract topics.	• Produces extended stretches of language with flexibility and ease and very little hesitation. • Contributions are relevant, coherent, varied and detailed. • Makes full and effective use of a wide range of cohesive devices and discourse markers.	• Is intelligible. • Phonological features are used effectively to convey and enhance meaning.	• Interacts with ease by skilfully interweaving his/her contributions into the conversation. • Widens the scope of the interaction and develops it fully and effectively towards a negotiated outcome.
C2	• Maintains control of a wide range of grammatical forms.	• Uses a wide range of appropriate vocabulary to give and exchange views on unfamiliar and abstract topics.	• Produces extended stretches of language with ease and with very little hesitation. • Contributions are relevant, coherent and varied. • Uses a wide range of cohesive devices and discourse markers.	• Is intelligible. • Intonation is appropriate. • Sentence and word stress is accurately placed. • Individual sounds are articulated clearly.	• Interacts with ease, linking contributions to those of other speakers. • Widens the scope of the interaction and negotiates towards an outcome.
C1	• Shows a good degree of control of a range of simple and some complex grammatical forms.	• Uses a range of appropriate vocabulary to give and exchange views on familiar and unfamiliar topics.	• Produces extended stretches of language with very little hesitation. • Contributions are relevant and there is a clear organisation of ideas. • Uses a range of cohesive devices and discourse markers.	• Is intelligible. • Intonation is appropriate. • Sentence and word stress is accurately placed. • Individual sounds are articulated clearly.	• Initiates and responds appropriately, linking contributions to those of other speakers. • Maintains and develops the interaction and negotiates towards an outcome.
	Grammar and Vocabulary				
B2	• Shows a good degree of control of simple grammatical forms, and attempts some complex grammatical forms. • Uses appropriate vocabulary to give and exchange views, on a range of familiar topics.		• Produces extended stretches of language despite some hesitation. • Contributions are relevant and there is very little repetition. • Uses a range of cohesive devices.	• Is intelligible. • Intonation is generally appropriate. • Sentence and word stress is generally accurately placed. • Individual sounds are generally articulated clearly.	• Initiates and responds appropriately. • Maintains and develops the interaction and negotiates towards an outcome with very little support.
B1	• Shows a good degree of control of simple grammatical forms. • Uses a range of appropriate vocabulary when talking about familiar topics.		• Produces responses which are extended beyond short phrases, despite hesitation. • Contributions are mostly relevant, but there may be some repetition. • Uses basic cohesive devices.	• Is mostly intelligible, and has some control of phonological features at both utterance and word levels.	• Initiates and responds appropriately. • Keeps the interaction going with very little prompting and support.
A2	• Shows sufficient control of simple grammatical forms. • Uses appropriate vocabulary to talk about everyday situations.			• Is mostly intelligible, despite limited control of phonological features.	• Maintains simple exchanges, despite some difficulty. • Requires prompting and support.
A1	• Shows only limited control of a few grammatical forms. • Uses a vocabulary of isolated words and phrases.			• Has very limited control of phonological features and is often unintelligible.	• Has considerable difficulty maintaining simple exchanges. • Requires additional prompting and support.

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

Appendix IB

Speaking assessment

Glossary of terms

1. GENERAL	
Conveying basic meaning	**Conveying basic meaning:** the ability of candidates to get their message across to their listeners, despite possible inaccuracies in the structure and/or delivery of the message.
Situations and topics	**Everyday situations:** situations that candidates come across in their everyday lives, e.g. having a meal, asking for information, shopping, going out with friends or family, travelling to school or work, taking part in leisure activities. A *Cambridge English: Key (KET)* task that requires candidates to exchange details about a store's opening hours exemplifies an everyday situation. **Familiar topics:** topics about which candidates can be expected to have some knowledge or personal experience. *Cambridge English: First (FCE)* tasks that require candidates to talk about what people like to do on holiday, or what it is like to do different jobs, exemplify familiar topics. **Unfamiliar topics:** topics which candidates would not be expected to have much personal experience of. *Cambridge English: Advanced (CAE)* tasks that require candidates to speculate about whether people in the world today only care about themselves, or the kinds of problems that having a lot of money can cause, exemplify unfamiliar topics. **Abstract topics:** topics which include ideas rather than concrete situations or events. *Cambridge English: Proficiency (CPE)* tasks that require candidates to discuss how far the development of our civilisation has been affected by chance discoveries or events, or the impact of writing on society, exemplify abstract topics.
Utterance	**Utterance:** people generally write in sentences and they speak in utterances. An utterance may be as short as a word or phrase, or a longer stretch of language.

2. GRAMMAR AND VOCABULARY	
Appropriacy of vocabulary	**Appropriacy of vocabulary:** the use of words and phrases that fit the context of the given task. For example, in the utterance *I'm very sensible to noise*, the word *sensible* is inappropriate as the word should be *sensitive*. Another example would be *Today's big snow makes getting around the city difficult*. The phrase *getting around* is well suited to this situation. However, *big snow* is inappropriate as *big* and *snow* are not used together. *Heavy snow* would be appropriate.
Flexibility	**Flexibility:** the ability of candidates to adapt the language they use in order to give emphasis, to differentiate according to the context, and to eliminate ambiguity. Examples of this would be reformulating and paraphrasing ideas.
Grammatical control	**Grammatical control:** the ability to consistently use grammar accurately and appropriately to convey intended meaning. Where language specifications are provided at lower levels (as in *Cambridge English: Key (KET)* and *Cambridge English: Preliminary (PET)*), candidates may have control of only the simplest exponents of the listed forms. **Attempts at control:** sporadic and inconsistent use of accurate and appropriate grammatical forms. For example, the inconsistent use of one form in terms of structure or meaning, the production of one part of a complex form incorrectly or the use of some complex forms correctly and some incorrectly. Spoken language often involves false starts, incomplete utterances, ellipsis and reformulation. Where communication is achieved, such features are not penalised.

2. GRAMMAR AND VOCABULARY (cont.)	
Grammatical forms	**Simple grammatical forms:** words, phrases, basic tenses and simple clauses. **Complex grammatical forms:** longer and more complex utterances, e.g. noun clauses, relative and adverb clauses, subordination, passive forms, infinitives, verb patterns, modal forms and tense contrasts.
Range	**Range:** the variety of words and grammatical forms a candidate uses. At higher levels, candidates will make increasing use of a greater variety of words, fixed phrases, collocations and grammatical forms.

3. DISCOURSE MANAGEMENT	
Coherence and cohesion	**Coherence and cohesion** are difficult to separate in discourse. Broadly speaking, coherence refers to a clear and logical stretch of speech which can be easily followed by a listener. Cohesion refers to a stretch of speech which is unified and structurally organised. Coherence and cohesion can be achieved in a variety of ways, including with the use of cohesive devices, related vocabulary, grammar and discourse markers. **Cohesive devices:** words or phrases which indicate relationships between utterances, e.g. addition (*and, in addition, moreover*); consequence (*so, therefore, as a result*); order of information (*first, second, next, finally*). At higher levels, candidates should be able to provide cohesion not just with basic cohesive devices (*e.g. and, but, or, then, finally*) but also with more sophisticated devices (*e.g. therefore, moreover, as a result, in addition, however, on the other hand*). **Related vocabulary:** the use of several items from the same lexical set, e.g. *train, station, platform, carriage*; or *study, learn, revise*. **Grammatical devices:** essentially the use of reference pronouns (*e.g. it, this, one*) and articles (*e.g. There are two women in the picture. The one on the right ...*). **Discourse markers:** words or phrases which are primarily used in spoken language to add meaning to the interaction, e.g. *you know, you see, actually, basically, I mean, well, anyway, like*.
Extent/extended stretches of language	**Extent/extended stretches of language:** the amount of language produced by a candidate which should be appropriate to the task. Long turn tasks require longer stretches of language, whereas tasks which involve discussion or answering questions could require shorter and extended responses.
Relevance	**Relevance:** a contribution that is related to the task and not about something completely different.
Repetition	**Repetition:** repeating the same idea instead of introducing new ideas to develop the topic.

4. PRONUNCIATION	
Intelligible	**Intelligible:** a contribution which can generally be understood by a non-EFL/ESOL specialist, even if the speaker has a strong or unfamiliar accent.
Phonological features	**Phonological features** include the pronunciation of individual sounds, word and sentence stress and intonation. **Individual sounds are:** • Pronounced vowels, e.g. the /æ/ in cat or the /e/ in bed • Diphthongs, when two vowels are rolled together to produce one sound, e.g. the /əʊ/ in host or the /eɪ/ in hate • Consonants, e.g. the /k/ in cut or the /f/ in fish. **Stress:** the emphasis laid on a syllable or word. Words of two or more syllables have one syllable which stands out from the rest because it is pronounced more loudly and clearly, and is longer than the others, e.g. imPORtant. Word stress can also distinguish between words, e.g. proTEST vs PROtest. In sentences, stress can be used to indicate important meaning, e.g. *WHY is that one important?* versus *Why is THAT one important?* **Intonation:** The way the voice rises and falls, e.g. to convey the speaker's mood, to support meaning or to indicate new information.

5. INTERACTIVE COMMUNICATION	
Development of the interaction	**Development of the interaction:** actively developing the conversation, e.g. by saying more than the minimum in response to the written or visual stimulus, or to something the other candidate/ interlocutor has said, or by proactively involving the other candidate with a suggestion or question about further developing the topic (e.g. *What about bringing a camera for the holiday?* or *Why's that?*).
Initiating and Responding	**Initiating:** starting a new turn by introducing a new idea or a new development of the current topic. **Responding:** replying or reacting to what the other candidate or the interlocutor has said.
Prompting and Supporting	**Prompting:** instances when the interlocutor repeats, or uses a backup prompt or gesture in order to get the candidate to respond or make a further contribution. **Supporting:** instances when one candidate helps another candidate, e.g. by providing a word they are looking for during a discussion activity, or helping them develop an idea.
Turn and Simple exchange	**Turn:** everything a person says before someone else speaks. **Simple exchange:** a brief interaction which typically involves two turns in the form of an initiation and a response, e.g. question-answer, suggestion-agreement.

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

Appendix 1C Holistic Rating Scale used by the Interlocutor

Cambridge English: First Speaking Examiners use a more detailed version of the following assessment scales, extracted from the overall Speaking scales on page 83:

B2	Grammar and Vocabulary	Discourse Management	Pronunciation	Interactive Communication
5	Shows a good degree of control of a range of simple and some complex grammatical forms. Uses a range of appropriate vocabulary to give and exchange views on a wide range of familiar topics.	Produces extended stretches of language with very little hesitation. Contributions are relevant and there is a clear organisation of ideas. Uses a range of cohesive devices and discourse markers.	Is intelligible. Intonation is appropriate. Sentence and word stress is accurately placed. Individual sounds are articulated clearly.	Initiates and responds appropriately, linking contributions to those of other speakers. Maintains and develops the interaction and negotiates towards an outcome.
4	*Performance shares features of Bands 3 and 5.*			
3	Shows a good degree of control of simple grammatical forms, and attempts some complex grammatical forms. Uses a range of appropriate vocabulary to give and exchange views on a range of familiar topics.	Produces extended stretches of language despite some hesitation. Contributions are relevant and there is very little repetition. Uses a range of cohesive devices.	Is intelligible. Intonation is generally appropriate. Sentence and word stress is generally accurately placed. Individual sounds are generally articulated clearly.	Initiates and responds appropriately. Maintains and develops the interaction and negotiates towards an outcome with very little support.
2	*Performance shares features of Bands 1 and 3.*			
1	Shows a good degree of control of simple grammatical forms. Uses a range of appropriate vocabulary when talking about everyday situations.	Produces responses which are extended beyond short phrases, despite hesitation. Contributions are mostly relevant, despite some repetition. Uses basic cohesive devices.	Is mostly intelligible, and has some control of phonological features at both utterance and word levels.	Initiates and responds appropriately. Keeps the interaction going with very little prompting and support.
0	*Performance below Band 1.*			

B2	Global Achievement
5	Handles communication on a range of familiar topics, with very little hesitation. Uses accurate and appropriate linguistic resources to express ideas and produce extended discourse that is generally coherent.
4	*Performance shares features of Bands 3 and 5.*
3	Handles communication on familiar topics, despite some hesitation. Organises extended discourse but occasionally produces utterances that lack coherence, and some inaccuracies and inappropriate usage occur.
2	*Performance shares features of Bands 1 and 3.*
1	Handles communication in everyday situations, despite hesitation. Constructs longer utterances but is not able to use complex language except in well-rehearsed utterances.
0	*Performance below Band 1.*

Appendix II

BULATS
Business Language Testing Service

BULATS Online Speaking

Assessment Criteria: Parts 1, 3, 4 & 5

Public version

BAND	GLOBAL DESCRIPTORS
6 (C2) • task achievement • coherence / discourse management • language resource • pronunciation • hesitation / extent	**Fully operational command of the spoken language** • Achieves the task effectively; responses are consistently appropriate. • Able to express both simple and complex ideas with ease; coherent extended discourse. • Consistently, displays wide range and accurate use of grammar and vocabulary. • Pronunciation is easy to understand; stress, rhythm and intonation are used to express meaning effectively. • Responds promptly with only natural hesitation; makes effective use of the allowed response time.
5 (C1) • task achievement • coherence / discourse management • language resource • pronunciation • hesitation / extent	**Good operational command of the spoken language** • Achieves the task well; responses are generally appropriate. • Able to express simple and complex ideas; generally extends discourse coherently. • Generally, displays wide range and accurate use of grammar and vocabulary. • Pronunciation is easy to understand; stress, rhythm and intonation are used to express meaning well. • Generally responds promptly, with only natural hesitation; generally makes good use of the allowed response time.
4 (B2) • task achievement • coherence / discourse management • language resource • pronunciation • hesitation / extent	**Generally effective command of the spoken language** • Achieves the task adequately; most responses are appropriate but a few may be inappropriate or ambiguous (possibly due to incomprehension of input). • Able to express simple ideas and makes some attempt to express complex ideas; mostly coherent, with some extended discourse. • There is an adequate range of grammar and vocabulary which is sufficiently accurate to deal with the tasks. • Pronunciation can generally be understood; stress, rhythm and intonation are used to express meaning adequately. • May be some hesitation while searching for language; generally makes adequate use of the allowed response time.
3 (B1) • task achievement • coherence / discourse management • language resource • pronunciation • hesitation / extent	**Limited but effective command of the spoken language** • Achieves most of the task, in a limited way; some responses may be inappropriate, ambiguous or not attempted (possibly due to incomprehension of input). • Able to express simple ideas; little extended discourse; some incoherence. • The range of grammar and vocabulary used is sufficient to complete tasks in a limited way. Some language in simple utterances is accurate but basic inaccuracies may impede communication of ideas and achievement of the tasks. • Pronunciation can generally be understood but L1 features may cause strain; an attempt is made to use aspects of stress, rhythm and intonation to express meaning. • Hesitation may demand patience of the listener; use of the allowed response time may not always be adequate.
2 (A2) • task achievement • coherence / discourse management • language resource • pronunciation • hesitation / extent	**Basic command of the spoken language** • Achieves only simplest part of the task (i.e. responding to simple prompts) in a very limited way; many responses may be inappropriate, ambiguous or not attempted (possibly due to incomprehension of input). • No extended discourse • The range of language is sufficient to respond to simple prompts but not to complete complex tasks. Some utterances (single words or short phrases) may be accurate but inaccuracies in grammar and vocabulary limit achievement of the tasks and restrict coherence and communication of ideas. • Pronunciation of single words may be intelligible but L1 features may make understanding difficult; little attempt is made to use aspects of stress, rhythm and intonation to express meaning. • Hesitation is excessive; use of the allowed response time is adequate on only a few occasions.
1 (A1) • task achievement • coherence / discourse management • language resource • pronunciation • hesitation / extent	**Minimal command of the spoken language** • May achieve a few of the simplest parts of the task (i.e. responding to simple prompts) in a very limited way; most responses may be inappropriate, ambiguous or not attempted (possibly due to incomprehension of input). • Utterances may be limited to single words. • The range of language is limited and inadequate to complete the tasks. Some accurate language but frequent inaccuracies may mean the message is not communicated. • Pronunciation of single words may be intelligible but L1 features may cause excessive strain to a listener; no attempt is made to use aspects of stress, rhythm and intonation to express meaning. • Hesitation is excessive; use of the allowed response time is generally inadequate.
0	Throughout the task, responses are not attempted, OR consistently no meaning is conveyed, OR responses are consistently unrelated to the rubric.

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

Appendix III

BULATS
Business Language Testing Service

BULATS Online Speaking

Assessment Criteria: Part 2 'Reading Aloud' task

Public version

BAND	DESCRIPTORS for Reading Aloud
6 (C2) • overall intelligibility • individual sounds • stress etc	• Pronunciation is easy to understand and meaning is conveyed effectively. • Individual sounds are clear and unambiguous. • Stress, rhythm and intonation are consistently used appropriately so that meaning is expressed effectively.
5 (C1) • overall intelligibility • individual sounds • stress etc	• Pronunciation is easy to understand and meaning is conveyed well. • Individual sounds are generally clear and unambiguous. • Stress, rhythm and intonation are generally used appropriately so that meaning is expressed well.
4 (B2) • overall intelligibility • individual sounds • stress etc	• Pronunciation can generally be understood and meaning is conveyed adequately. • Individual sounds are generally clear although there may be occasional difficulty for the listener. • Stress, rhythm and intonation are used to express meaning adequately.
3 (B1) • overall intelligibility • individual sounds • stress etc	• Pronunciation can generally be understood but L1 features may cause strain; meaning is conveyed but there may be some ambiguity. • Many individual sounds are clear but some may cause difficulty for the listener. • An attempt is made to use aspects of stress, rhythm and intonation to express meaning.
2 (A2) • overall intelligibility • individual sounds • stress etc	• Pronunciation of single words may be intelligible but L1 features may make understanding difficult and some meaning may be distorted. • Inaccuracies in the pronunciation of individual sounds may cause strain for the listener and may impede communication of meaning. • Little attempt is made to use aspects of stress, rhythm and intonation to express meaning.
1 (A1) • overall intelligibility • individual sounds • stress etc	• Pronunciation of single words may be intelligible but L1 features may cause excessive strain to a listener and meaning may be seriously distorted. • Serious inaccuracies in the pronunciation of individual sounds may make speech unintelligible. • No attempt is made to use aspects of stress, rhythm and intonation to express meaning.
0	Responses not attempted OR not enough language to assess.

Appendix IV

Appendix 3 Operational checklist (used in Phase 3)

Informational functions	
Providing personal information	• Give information on present circumstances • Give information on past experiences • Give information on future plans
Expressing opinions	Express opinions
Elaborating	Elaborate on, or modify an opinion
Justifying opinions	Express reasons for assertions s/he had made
Comparing	Compare things/people/events
Speculating	Speculate
Staging	Separate out or interpret the parts of an issue
Describing	• Describe a sequence of events • Describe a scene
Summarizing	Summarize what s/he has said
Suggesting	Suggest a particular idea
Expressing preferences	Express preferences
Interactional functions	
Agreeing	Agree with an assertion made by another speaker (apart from 'yeah' or nonverbal)
Disagreeing	Disagree with what another speaker says (apart from 'no' or nonverbal)
Modifying	Modify arguments or comments made by other speaker or by the test-taker in response to another speaker
Asking for opinions	Ask for opinions
Persuading	Attempt to persuade another person
Asking for information	Ask for information
Conversational repair	Repair breakdowns in interaction
Negotiating meaning	• Check understanding • Indicate understanding of point made by partner • Establish common ground/purpose or strategy • Ask for clarification when an utterance is misheard or misinterpreted • Correct an utterance made by other speaker which is perceived to be incorrect or inaccurate • Respond to requests for clarification
Managing interaction	
Initiating	Start any interactions
Changing	Take the opportunity to change the topic
Reciprocating	Share the responsibility for developing the interaction
Deciding	Come to a decision