Cambridge Assessment English

Linguaskill

Building a validity argument for the Speaking test

Jing Xu, Mark Brenchley, Edmund Jones, Annabelle Pinnington, Trevor Benjamin, Kate Knill, Gaelle Seal-Coon, Martin Robinson and Ardeshir Geranpayeh

Linguaskill

Contents

Summary 3

Introduction 4

1. Test purpose 5

2. Target language use domain 5

3. Test description 5

4. Marking of speaking performance 6

5. Framework of test validation 8

6. Validity evidence for the Linguaskill Speaking test 11

References 17

Summary

The Linguaskill Speaking test is a computer-based oral English test that is enhanced by auto-marking technology. It takes a hybrid approach to marking which combines the strengths and benefits of artificial intelligence with those of the decision-making of experienced human markers. The test is appealing to institutional users who may need to assess a large number of English language learners in a short time frame. Individual learners may also find the browser-based speaking test highly accessible in that it can be taken at home on any Windows computer with a high-speed internet connection. The test results of Linguaskill are reported within 48 hours thanks to auto-marking technology.

This paper presents a validity argument for the Linguaskill Speaking test by weaving together a narrative about the research evidence that has been collected to support the intended interpretations and uses of the test scores. We begin by describing the purpose, the target language use (TLU) domain and the format of the test. Then we unveil the design of the auto-marker, the training programme of human examiners and the hybrid marking model used in the test. In what follows we delineate the structure of the Linguaskill validity argument and explain why each element in it is essential for arguing for test validity. Finally, we present research evidence related to three elements in the validity argument, namely *Test Content*, *Marking of Responses*, and *Interpretation of Test Results*. It is notable that test validation is a cumulative process. Further research is being carried out to gather additional evidence to support the validity argument.

The evidence presented in this paper supports the following validity claims about the test.

- **Test Content:** The speaking topics which cover daily routines, dialogues at social activities, exchanges at workplaces and telecommunications are overall a good representation of the communicative situations that the candidates will likely encounter in the real world.
- **Test Content:** The speaking topics are interesting, neither too easy nor too difficult.
- **Test Content:** Candidates generally feel comfortable with speaking to a computer.
- **Marking of Responses:** The reliability of human marking is satisfactory.
- **Marking of Responses:** With hybrid marking in place, the auto-marker achieved 95.6% exact agreement and 100% adjacent agreement on Common European Framework of Reference (CEFR) grades with human examiners.
- **Marking of Responses:** The auto-marker is capable of detecting suspicious non-English speech and escalating it to human examiners for verification.
- **Interpretation of Test Results:** Standard-setting exercises are conducted regularly to establish the link between the test and the CEFR. For this reason, the test results can be interpreted confidently based on the CEFR.
- **Interpretation of Test Results:** Confirmatory factor analysis suggests that a single, overarching speaking construct is assessed by the test.

Introduction

The Linguaskill Speaking test is a computer-based oral English test that is enhanced by auto-marking technology. In contrast to other automated speaking assessments, Linguaskill Speaking takes a hybrid approach to marking, which means its test responses are marked by a combination of human examiners and auto-marking technology. If the computer indicates low confidence in marking a response, the response is escalated to human marking. This hybrid model aims to address the challenges that fully automated assessment brings, by marrying the latest auto-marking technology with the decision-making of experienced human markers.

With the advancement of natural language processing, machine learning and speech recognition technologies, automated speaking assessment is growing in popularity. Compared with traditional face-to-face speaking exams, automated speaking assessment, which is delivered on a computer or mobile device, offers the benefits of much faster score reporting, simple test administration and on-demand testing.

Automated speaking assessment is particularly appealing to institutional users who may find large-scale administration of face-to-face speaking exams unfeasible. Individual learners may see increased accessibility in automated assessment because, with a remote proctoring solution in place, the speaking test can be taken even at home. However, automated assessment also brings about challenges and problems that are not typically associated with human marking, such as scepticism about construct coverage and susceptibility to candidate cheating (Chun 2006, Fan 2014, Xi 2010, Xu 2015).

This paper presents a validity argument for the Linguaskill Speaking test. A validity argument provides an overall evaluation of the intended interpretations and uses of test scores by conducting coherent analysis on various strands of research evidence either for or against the proposed interpretations and uses (Cronbach 1988, Kane 2013). We begin by describing the test specifications including the intended test purpose, target language use domain and test format. Then, we introduce the hybrid marking model applied to the oral assessment in which marking tasks are shared by an auto-marker and human examiners. In what follows, we present a clear validation framework which lays out critical validity considerations at different stages of a testing cycle. In the remainder of the paper, we present validity evidence that has been collected based on this framework.

1. Test purpose

The Linguaskill Speaking test assesses candidates' oral English proficiency for everyday communication. It can be taken on its own or in conjunction with the other Linguaskill modules of Reading and Listening, and Writing. Linguaskill aims to provide fast, reliable and clearly interpretable results based on the Common European Framework of Reference for Languages (CEFR), a widely recognised standard for describing the progression of language learning and acquisition (Council of Europe 2001, 2018), as well as more granular scores based on the Cambridge English Scale. Intended uses of Linguaskill include a) measuring a candidate's level of English for placement, progression, or graduation at education institutions and b) measuring a candidate's level of English for job or development opportunities at companies. The target candidates of Linguaskill are English language learners over the age of 16 years.

2. Target language use domain

A target language use (TLU) domain is a hypothetical description of the situations or contexts in which candidates need to be able to use the language outside the test. By delineating the scope of this domain and identifying the key characteristics of language use in it, the test developer is able to design tasks that mimic these language use activities. Then, candidates' test-taking behaviours can be viewed as a sample of their predicted language performance in the TLU domain.

As Linguaskill is designed to serve multiple test purposes (see Section 1), its TLU domain has to be fairly broad, covering a wide range of situations and tasks of English language use in both daily-life and workplace settings. As the contexts of communication in this domain are heterogeneous, it is important for the test developer to identify and describe some critical TLU tasks and ensure that they are represented in test content (see a further discussion in Section 5.1).

The critical TLU tasks selected for Linguaskill generally fall into four categories. They are daily routines (e.g., discussing leisure-time habits, giving preferences), dialogues at social activities (e.g., describing a situation or issue, recounting news or personal experience), exchanges at workplaces (e.g., raising a problem or issue, reporting on data), and telecommunications (e.g., requesting information and leaving a telephone message).

3. Test description

The Linguaskill Speaking test is browser-based so candidates can sit the test on any Windows computer[1] with a high-speed internet connection in an invigilated setting. The test is remotely proctored if a candidate chooses to take it at home. Questions are presented to the candidate through the computer screen and headphones, and their responses are recorded and remotely assessed by either computer algorithms or examiners (see Section 4). The test is multi-level, meaning that it is designed to elicit and assess oral performance of multiple proficiency levels based on the CEFR, including below A1, A1, A2, B1, B2, and C1 and above. The test results are reported within 48 hours.

The Linguaskill Speaking test has five parts: Interview, Reading Aloud, Presentation, Presentation with Visual Information, and Communication Activity. All parts are weighted equally and focus on different aspects of speaking ability. The format, testing aim and evaluation criteria of the five parts are presented below and summarised in Table 1.

3.1 Interview

a. Format

In the interview task, the candidate answers eight questions about themselves. The first four questions are standard in all tests and candidates have 10 seconds to answer each question. Questions 5–8 vary according to each test version and are likely to ask the candidate simple personal questions relating to habits, experiences, or tastes. Candidates have 20 seconds to answer each of these questions.

Table 1. An overview of the Linguaskill Speaking tasks

Part	Task	Description	Length of response(s)	Preparation time	Marks
1	Interview	The candidate answers eight questions about themselves.	4 x 10 secs and 4 x 20 secs	none	20%
2	Reading Aloud	The candidate reads aloud eight sentences.	8 x 10 secs	none	20%
3	Presentation	The candidate speaks on a given topic.	1 minute	40 secs	20%
4	Presentation with Visual Information	The candidate gives a presentation based on the graphic information given.	1 minute	1 minute	20%
5	Communication Activity	The candidate gives opinions on five questions related to a scenario.	5 x 20 secs	40 secs	20%

[1] At this point, sitting Linguaskill on a Mac computer is not supported. It is suggested that candidates use Google Chrome or Mozilla Firefox to take the test on a PC.

b. Testing aim
As well as introducing the candidate to the computer format of the test, the focus of this test part is to assess the candidate's ability to answer personal questions and to give lower-proficiency candidates a more accessible and achievable task.

c. Evaluation criteria
Candidates are assessed on their linguistic output in terms of pronunciation and fluency, and language resource.

3.2 Reading Aloud

a. Format
In the Reading Aloud task, the candidate is required to read aloud eight sentences. They have 10 seconds to read aloud each sentence. Sentences are of the kind that candidates may have to read aloud in real-world situations and are presented in increasing level of difficulty, covering a wide range of phonological features and syntactic structures.

b. Testing aim
The focus of this test part is to assess the candidate's ability to transform the written form of the language into speech and to handle elements of pronunciation at sentence level.

c. Evaluation criteria
Candidates are assessed based on phonological criteria, including their overall intelligibility, their ability to produce individual sounds, as well as their stress, rhythm and intonation.

3.3 Presentation

a. Format
In the Presentation task, the candidate is required to speak for 1 minute on a given topic. There is no choice of topic. A preparation time of 40 seconds is given before a candidate records their response.

b. Testing aim
The focus of this test part is to assess the candidate's ability to deliver a long turn. As well as a description of a situation or issue, the candidate is encouraged to state and/or justify an opinion through bulleted prompts.

c. Evaluation criteria
Candidates are assessed on their linguistic output in terms of pronunciation and fluency, language resource and discourse management. Additionally, the marking takes into account the candidate's ability to complete the task appropriately in accordance with the rubric and instructions.

3.4 Presentation with Visual Information

a. Format
In the Presentation with Visual Information task, the candidate is required to talk for 1 minute about information presented to them in visual form. A preparation time of 1 minute is given before a candidate records their response. The candidate is asked to present the information within a specific context, such as leaving a voicemail for a friend or giving a presentation in class.

b. Testing aim
The focus of this test part is to assess the candidate's ability to deliver a long turn which involves the interpretation of very simple visual information and providing a recommendation, explanation or suggestion.

c. Evaluation criteria
Candidates are assessed on their linguistic output in terms of pronunciation and fluency, language resource and discourse management. Additionally, the marking takes into account the candidate's ability to complete the task appropriately in accordance with the rubric and instructions.

3.5 Communication Activity

a. Format
In the Communication Activity task, the candidate is required to answer five questions related to a scenario. A preparation time of 40 seconds is given before the candidate hears the first question. Each question has a 20-second response window and asks the candidate to provide an opinion, speculate a hypothesis, or make an evaluation.

b. Testing aim
The focus of this test part is to assess the candidate's ability to express opinions and ideas on a given topic in response to an aural prompt. It is an opportunity for higher-level candidates to demonstrate their higher-level skills.

c. Evaluation criteria
Candidates are assessed on their linguistic output in terms of pronunciation and fluency, language resource and discourse management. Additionally, the marking takes into account the candidate's ability to complete the task appropriately in accordance with the rubric and instructions.

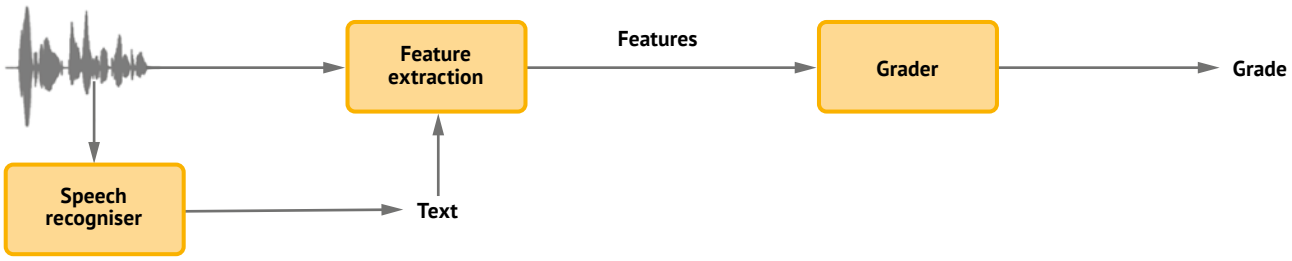
4. Marking of speaking performance

The Linguaskill Speaking test adopts a hybrid or *human-in-the-loop* marking model in which an auto-marker is used in live assessment, but with the involvement of human examiners. This section discusses the design of the auto-marker, examiner training and certification, and how hybrid marking is applied.

4.1 Auto-marker

An auto-marker is a set of computer algorithms designed to mark constructed test responses such as extended speaking and writing. Cambridge Assessment English (henceforth Cambridge English), in collaboration with research groups from the University of Cambridge, started to develop automated marking of spontaneous non-native English speech in 2012. The auto-marker used in the Linguaskill Speaking test is called the Custom Automated Speech Engine (CASE), which was developed by Enhanced Speech Technology Ltd building upon technology transferred from the

Figure 1. The architecture of the auto-marker (Knill et al 2018)



Institute for Automated Language Teaching and Assessment (ALTA), an interdisciplinary research centre of the University of Cambridge, using machine learning technologies.

CASE, as shown in Figure 1, consists of three major components: a speech recogniser, a feature extraction module, and a grader (Knill et al 2018, Wang et al 2018). The speech recogniser conducts Automated Speech Recognition (ASR), converting the audio signal of speech into a structured representation of the underlying word transcription. It was trained based on deep neural network models using learner speech supplied by Cambridge English and combined with crowd-sourced transcriptions (see the ASR2 system in Lu et al 2019). Feature extraction is about deriving features relevant to the speaking construct (e.g., fluency, pronunciation accuracy, vocabulary diversity) from both the audio signal and the structured word transcription as the basis for grading. Based on these features, the grader uses state-of-the-art machine learning models to return a distribution over scores from which feedback to the candidate such as the CEFR grade is derived. The training sample for the grader includes a large set of Linguaskill Speaking test responses produced by learners of various first languages and all CEFR levels as well as the marks awarded to these responses by examiners. In addition, the machine learning grader models used in the CASE have been selected to provide an uncertainty measure based on the similarity between the input and the training data (van Dalen, Knill and Gales 2015, Malinin, Ragni, Knill and Gales 2017). This uncertainty measure is a meaningful indicator of the reliability of the auto-marker score and is useful for identifying test responses that require human marking.

4.2 Examiners

All examiners for the Linguaskill Speaking test undergo a rigorous training programme in order to qualify (Figure 2). Prospective examiners must meet the minimum professional requirements, which include being educated to first degree level or equivalent, holding a recognised language teaching qualification, providing proof of substantial and relevant teaching experience within the last two years, and having suitable English language competency.

Approved applicants are provided with training materials through an online portal, which includes extensive documentation about the marking procedure and sample speaking responses with marks, along with detailed comments about performance and marking rationales. Applicants are guided through the material through the documentation.

Within 30 days of access the certification test must be successfully taken. Certification tests include a selection of speaking items previously marked by a pool of experienced and reliable examiners, with a statistically adjusted average score[2] as the final approved mark. Applicants must allocate a minimum of 80% of correct marks (within 0.5 of the approved mark). Two attempts are provided, with different versions of the test. Examiners who have failed both attempts have their access to the portal automatically revoked.

Once applicants have successfully passed the certification test the system identifies them as certificated and they are added to the marking pool and can start marking candidates. Once examiners start providing marks, they are continuously statistically monitored. Marking behaviour analysis is carried out to identify possible bias, consistency and non-compliant behaviour. Examiners who are flagged up statistically are investigated and removed from the marking pool if their behaviour is confirmed as unsatisfactory. Re-certification occurs every two years with new training and test material.

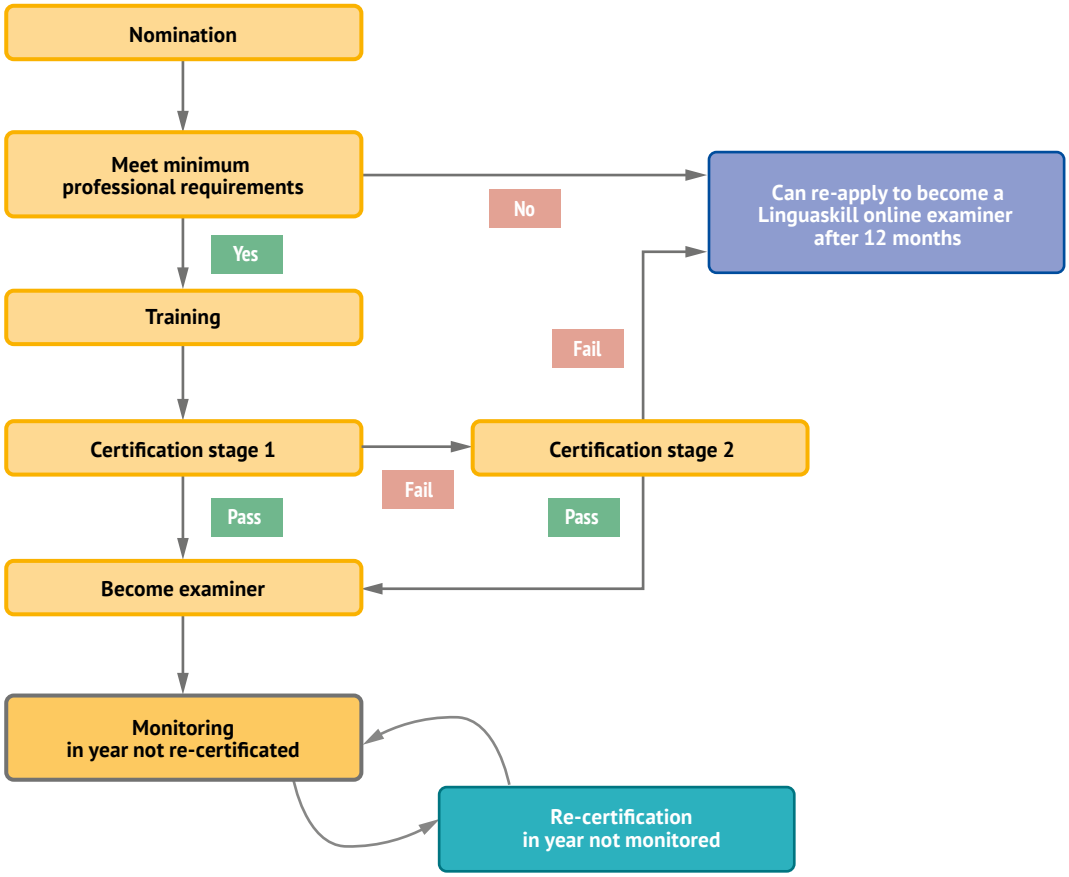
4.3 Hybrid marking

Hybrid marking aims to combine the strengths and benefits of artificial intelligence (AI) with those of human examiners. Computer marking is speedy and cost-effective but is only reliable when the responses being marked are close to the training sample of the AI system. Some impediments to auto-marker accuracy include poor audio quality, aberrant speaking behaviours and training sample underrepresentation. Poor audio quality is likely to significantly reduce ASR accuracy and affect the auto-marking performance. Learners may also be tempted to apply strategies to trick the marking system into giving them higher marks (Xi, Schmidgall and Wang 2016). Thus, it can be argued that human examiners play a key role as gatekeepers in preventing less reliable auto-marker scores being released to candidates.

The hybrid marking model is about using human examiner expertise to support and further develop the auto-marker. It is also based on the assumption that the computer can provide information to indicate its confidence in score prediction. When this confidence is low, the test response is flagged up and escalated to human examiners. In the

[2] We use fair average scores which are average scores adjusted for marker severity by multi-faceted Rasch analysis (Linacre 1989).

Figure 2. The procedure for certificating Linguaskill speaking examiners



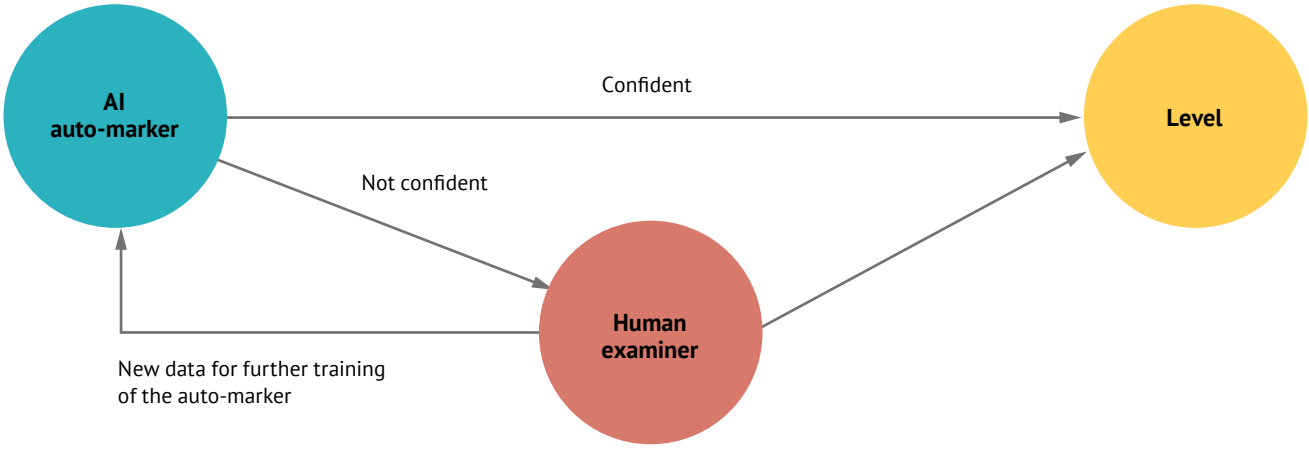
Cambridge English hybrid marking model (Figure 3), escalation to human marking is determined by setting thresholds on three features generated by the auto-marker in addition to the predicted score. They are the Assessment Quality score, Language Quality score and Audio Quality score. The *Assessment Quality* score is an uncertainty measure produced by the grader which suggests the amount of confidence the grader has in its score prediction (see Section 4.1). The *Language Quality* score is an ASR confidence score returned by the speech recogniser. It represents the system's confidence in the accuracy of its transcription, which in turn can be a useful proxy for identifying candidates who are not actually speaking English during the test (see Lu et al 2019). The *Audio Quality score* indicates the clarity of voice recording and is derived from three separate measures: dynamic ratio (differences in amplitude between loud and quiet parts of the audio), clipping (frames of audio that reach the maximum/minimum possible values and hence are distorted) and noise. It also incorporates a variety of other ASR errors linked to audio quality or the intermediate processes during the speech-to-text conversion. In addition, test responses with auto-marker scores falling below or above certain cut-off values are flagged for examiner marking. This is informed by our auto-marker evaluation suggesting that the auto-marker score tends to be less reliable on the lower and higher ends of the scoring scale. In the current hybrid marking model, a large proportion of test responses is marked by human examiners to ensure the quality of marking and provide marking data to further train the auto-marker. The proportion of human marking will gradually decrease with the enhancement of the auto-marker. The evaluation of the auto-marker and the hybrid marking model will be further discussed in Section 6.2.

5. Framework of test validation

Validity is the most fundamental issue of assessment. Validity is the degree to which evidence and theory support the interpretations of test results for intended test uses (AERA, APA and NCME 2014). Two common frameworks used for language test validation are the argument-based framework (Bachman and Palmer 2010, Kane 2013) and the socio-cognitive framework (Weir 2005). The former focuses on decomposing Messick's (1989) complex validity theory by structuring validity enquiry around practical arguments. The latter applies Messick's (1989) validity theory to language assessment, and similar to Messick, takes a cumulative approach to evidence collection. The Linguaskill validity argument, as shown in Figure 4, is constructed by integrating the two, in order to make validity claims and evaluate supporting evidence.

The Linguaskill validity argument consists of six parts: *Test Content*, *Response Processes*, *Marking of Responses*,

Figure 3. The Cambridge English hybrid marking model



Interpretation of Test Results, *Test Use* and *Test Impact*. They represent a sequence of activities that a typical testing cycle comprises, i.e., from test construction to the impact of testing on stakeholders. This section explains what these notions mean and how they can help guide the validation research for Linguaskill.

5.1 Test Content

A common understanding of test validity concerns the test itself or the instrument constructed to measure an ability. That is, is the test design of high quality and fit for purpose? This understanding is not incorrect, but it only addresses one facet of validity.

Traditionally, the aspect of validity concerning test content is called *content validity* (APA, AERA and NCME 1974) or *context validity* (Weir 2005). The idea is that the questions on a test should be relevant to intended test purposes and cover the critical knowledge and skills associated with such purposes. In language assessment, validity evidence for test content is usually gathered by expert review on the connection between test tasks and the TLU domain, which describes the situations or contexts the candidates are likely to encounter outside of the test. The aim of test content review is to ensure that the characteristics of the test tasks mirror or adequately represent the characteristics of language use activities in the TLU domain. This notion is also referred to as *authenticity* by some language testing researchers (e.g., Bachman and Palmer 1996) and is often considered as the basis of a validity argument (Bachman and Palmer 2010, Chapelle, Enright and Jamieson 2010).

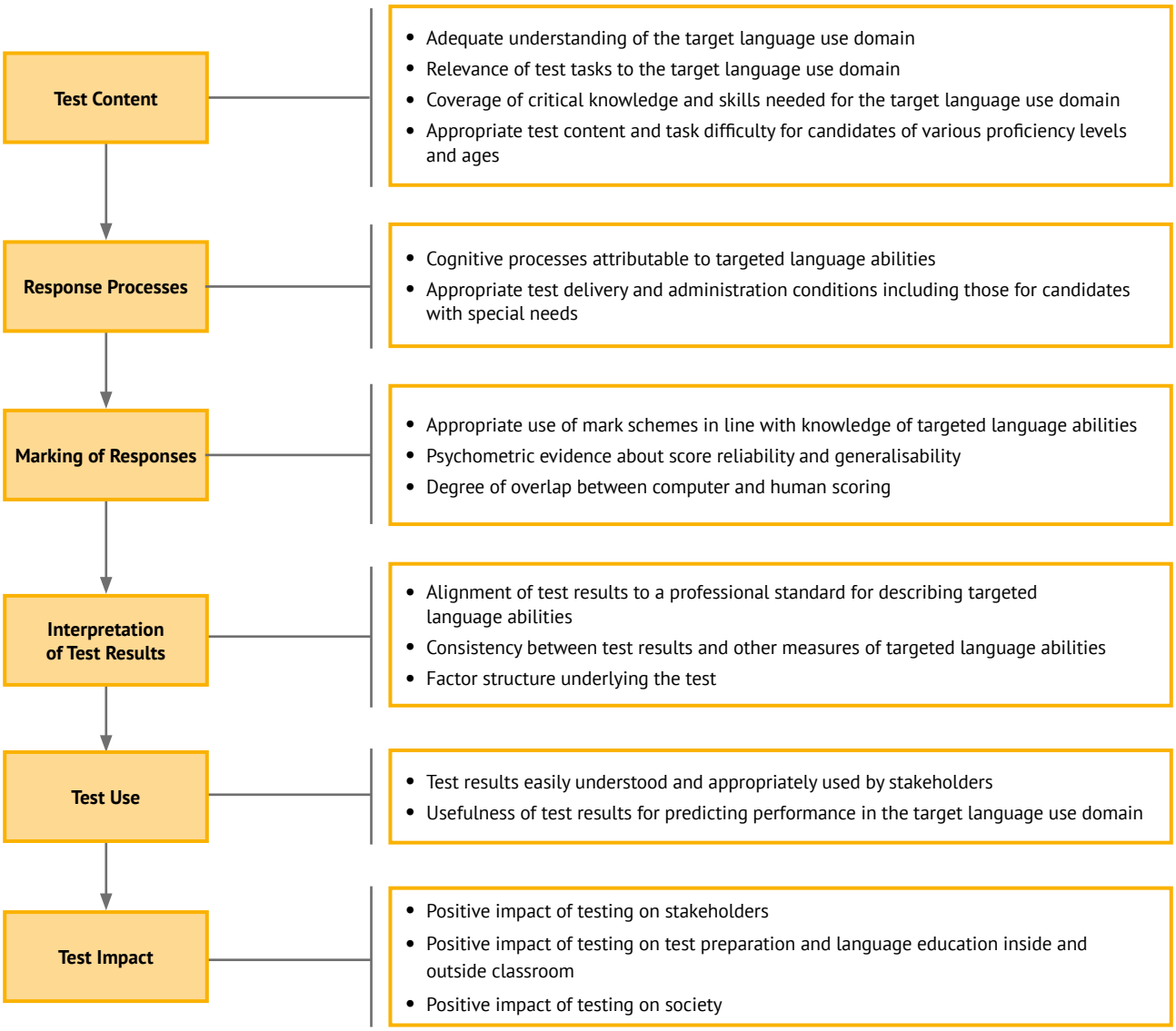
5.2 Response Processes

When candidates sit a language test, their language use is elicited and observed. Validity related to response processes concerns the elicitation of behaviours attributable to targeted language abilities. This ties to the original conception of *construct validity* regarding the traits or abilities being assessed by the test (Cronbach and Meehl 1955). Validity evidence for response processes includes coherence between test-taking behaviours/strategies and construct theories of language abilities, appropriate task delivery and administration, clear test instruction, and accommodation of candidates with special needs. Such evidence helps rule out an alternative interpretation of the test scores that factors other than targeted language ability had an effect on candidates' test performance (AERA, APA and NCME 2014).

5.3 Marking of Responses

Marking of test responses can be performed either by examiners or computer algorithms. When a test elicits constructed responses rather than selection from fixed multiple-choice options, mark schemes (also called scoring criteria) are needed to evaluate test performance. Validity evidence related to marking may come from mark scheme validation, analysis on human marking processes and rationales, and the investigation of consistency of marking. For example, it is expected that mark scheme development is informed by theories and research about targeted language abilities and that examiners assign credit to key aspects of language behaviours attributable to such abilities. Additionally, the marking processes, including the way scores are weighted or combined, should be justified and reflect the best approach to estimating targeted language abilities. It is also expected that the test score a candidate receives would provide a close estimate of the scores that he or she would have obtained on parallel forms of the same test or from any examiners randomly selected from the marker pool. This concept is commonly referred to as *reliability* (Haertel 2006) or *scoring validity* (Weir 2005). When a machine is used to predict human scores, the reliability of the automated scores must be presented in terms of machine agreement with human examiners or the deviation of machine scores from human scores. In addition, when constructed responses are marked by a computer, evidence about the degree of overlap between computer and human scoring criteria needs to be sought or it would be difficult to interpret computer scores and human scores in the same way (Xi 2010, Xu 2015).

Figure 4. The outline of the Linguaskill validity argument



5.4 Interpretation of Test Results

Test scores are simply numbers so meanings have to be assigned to them to make them useful for various purposes. This is at the heart of score interpretation. In most cases, the test developer provides test users with the suggested interpretations, in the form of Can Do statements of abilities associated with the scores, but this interpretation has to be backed up by theories of cognitive processes, language development, or second language acquisition (Weir 2005). Validity evidence for score interpretation can be obtained from standard-setting exercises that aim to align test scores to a theory-driven and/or research-based standard for describing language proficiency, such as the CEFR (Council of Europe 2001, 2018). Additionally, this evidence may be collected from concurrent studies that examine the relationship between test results and other measures of targeted language abilities. This is traditionally called *concurrent validity* (APA, AERA and NCME 1974). Validity evidence for score interpretation may also be collected from latent factor analysis which investigates the underlying factor structure of the test (e.g., Sawaki, Stricker and Oranje 2009). This piece of evidence is particularly relevant to integrated language assessment in which two or more language skills (e.g., listening and speaking) are assessed at the same time.

5.5 Test Use

Based on the test results, stakeholders, such as candidates, teachers, employers and admission officers, will likely take actions. For example, a candidate may decide to put in more effort to improve a particular skill; a teacher may tweak his or her lesson plans to meet students' learning needs; an employer may select a team among high-scoring candidates to expand overseas markets; a school admission officer may make acceptance decisions on applicants. Validity evidence related to test use is about the extent to which test results help stakeholders make informed decisions or take the right

actions. This evidence can be sought in the following two areas. First, suggested test use and meanings of test scores should be well understood by test users to avoid unintended score interpretations and uses. Second, test results should be useful for predicting future behaviours of interest such as job performance and academic performance that are related to language use. This use of validity in its predictive sense was called *predictive validity* and had been predominant before *construct validity* came into being in the 1950s (APA, AERA and NCME 1974).

5.6 Test Impact

The use of test scores will exert an impact on stakeholders in a range of teaching, learning and social contexts. For example, the way a high-stakes language test is designed is likely to influence how learners learn a language, how teachers teach a language, and even social values regarding language proficiency and fairness. This impact is also called *consequences*, *washback*, or *consequential validity* (Cheng 2014, Messick 1996, Weir 2005) and is an integral aspect of the concept of validity. Test impact is closely related to how a test is used. If Linguaskill were misused for unintended purposes, the test impact would probably be negative. The responsibility to ensure positive test impact is shared by both the test provider and test users.

6. Validity evidence for the Linguaskill Speaking test

This section presents the research evidence to support the use of the Linguaskill Speaking test for its intended purposes. Test validation is a cumulative and ongoing process (Messick 1989). Validity evidence is collected and refined over time as more data is collected, and as the Linguaskill test is relatively new, not all aspects of the validity argument have yet been fully documented. Extensive evidence has been obtained for *Test Content*, *Marking of Responses* and *Interpretation of Test Results*. Further research is being carried out to gather additional evidence for *Response Processes*, *Test Use* and *Test Impact* in specific countries and regions where the test is used.

6.1 Validity evidence for Test Content

Expert judgment is an essential element in attesting the relevance of the test tasks to the TLU domain (Messick 1989, p. 39). For Linguaskill, the judgment on content relevance and construct coverage (i.e., the skills assessed by the test) is made in test review meetings at the test development phase by a group of experts consisting of item writers, senior examiners and language testing researchers (Figure 5). The review of speaking items focuses on item difficulty, clarity of the prompt and instruction, authenticity of topics and situations, background knowledge needed to give a response, and the skills being assessed. Test items that fail the review are either discarded or revised before being reviewed once again. Those which pass the expert review are trialled with a large group of learners selected from the target test population before being included in a live test. Any problematic test items found by the trial are returned to the development phase.

Xu and Gallacher (2017) conducted a survey on 3,601 adult English language learners from 23 countries in a global trial of the Linguaskill Speaking test. The majority of the participants reported that the speaking tasks were similar to how they used English in the real world (65.3%) and that the speaking topics were closely related to their life (57.9%). In addition, approximately 70% of the participants agreed or strongly agreed that the Linguaskill Speaking test allowed them to demonstrate their English-speaking ability.

Analysis on the qualitative feedback received in the survey suggested that the speaking topics were interesting, neither too easy nor too difficult, and related to everyday life or work. For example, one participant related test questions to his daily language use situations:

> 'I find the topics relevant, not too easy nor difficult. I think that these topics are related to what normally happens in daily life. These are topics that most people learning English should master because they are what takes place in the real world.' (Participant ID 1742853)

Many participants reported the speaking test was not as stressful as they had anticipated. As one participant put it:

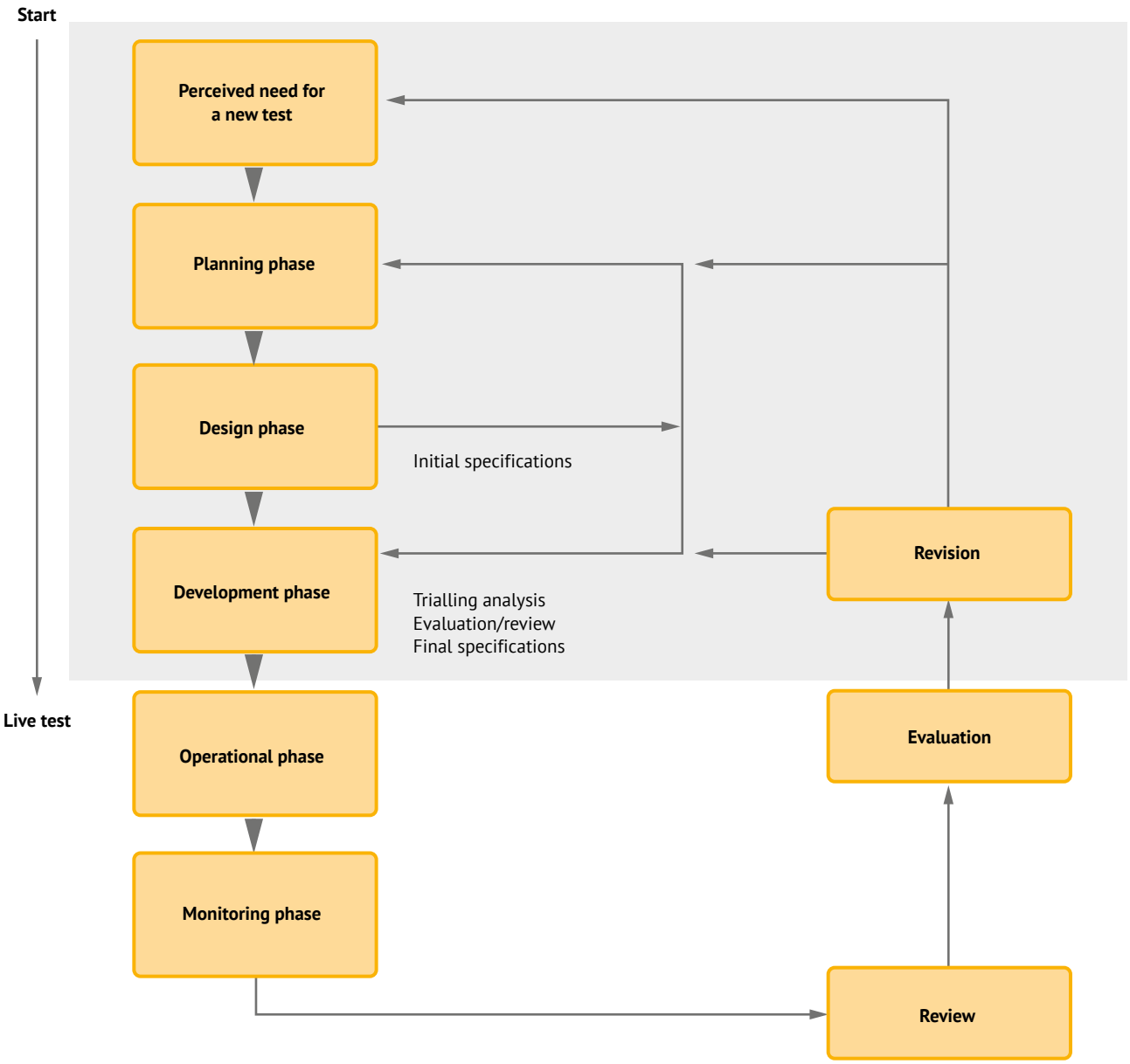
> 'I always feel worried in exams, but as I hear the questions, I felt more comfortable and relax. So they were easy for me.' (Participant ID 1721614)

Participants had mixed feelings about speaking to a computer. Some felt that talking to a computer was 'just like talking to a real person' (Participant ID 1750497) or even less stressful than talking to a speaking examiner (Participant ID 1741200). Others indicated that they were very used to interacting with a digital device since 'interaction through [a] mobile phone is quite popular nowadays' (Participant ID 1756618). A small proportion (19.3%) of participants still preferred to speak to a human interlocutor as they had expected exchanges of information (Participant ID 1646082) and seeing a human face (Participant ID 1714910) in the real-life oral communication.

In short, the quality-assurance process underpinning the Linguaskill content production and the findings from this large-scale trial study suggest that the content of the Linguaskill Speaking test is overall a good representation of the speaking tasks that candidates will likely encounter in the real world.

The ability to interact with an interlocutor (Brown 2003, Galaczi and Taylor 2018), which is typically assessed in a face-to-face interview, is not represented in the Linguaskill Speaking test. It is therefore not possible to interpret the Linguaskill Speaking test score as a direct measure of interactional competence. Nevertheless, monologic speaking performance may to some extent predict interactional speaking performance (Bernstein, Van Moere and Cheng 2010), and it

Figure 5. The test development cycle adopted by Linguaskill (Cambridge English 2016)



can be argued that the Linguaskill Speaking test is designed to cover a variety of communicative speaking functions.

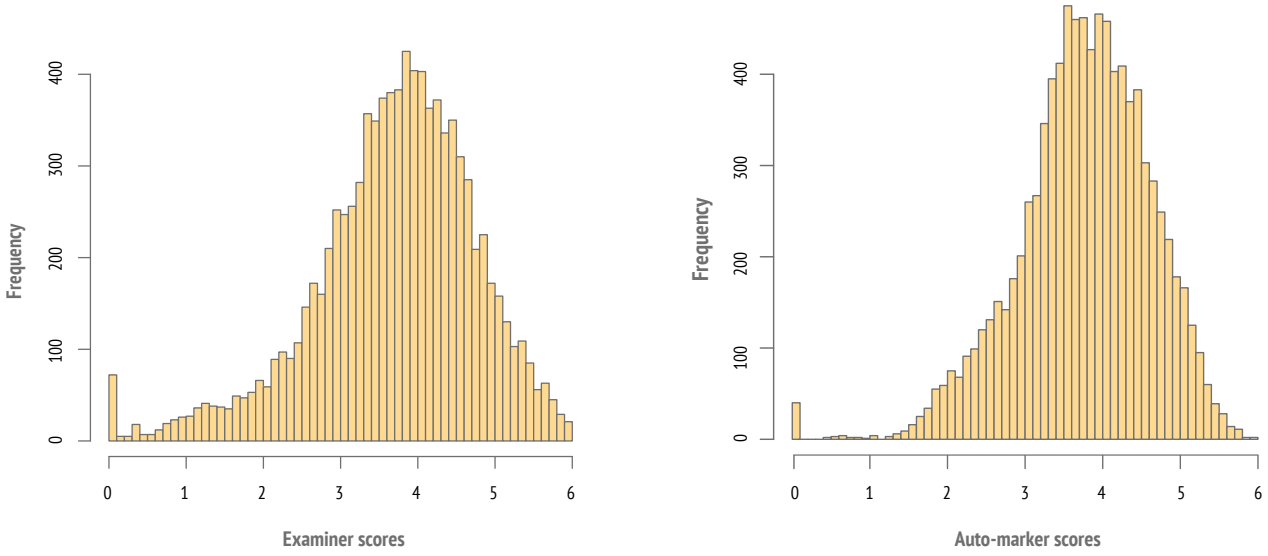
6.2 Validity evidence for Marking of Responses

Reliable marking of test responses serves as the basis for accurate estimation about candidates' targeted language abilities. A caveat associated with automated speaking assessment relates to the reliability of marking open-ended speech (Xu 2015). Before discussing the reliability of CASE, the speech auto-marker used in Linguaskill, we first report the reliability of examiner marking. This is because a) a large proportion of the Linguaskill Speaking tests are still marked by examiners and b) the auto-marker was trained on human-marked spoken data and thus cannot outperform the best examiners in the training sample.

6.2.1 Reliability of examiner marking

Xu and Gallacher (2017) conducted a study to investigate the reliability of human marking in the Linguaskill Speaking test. Five Linguaskill speaking examiners randomly selected from a larger pool were asked to mark a common dataset consisting of test responses produced by 60 candidates of various proficiency levels. In other words, each part of the test was marked by the same five examiners. Reliability of human marking on each test part and the whole test (which was the average of each part) was estimated using intraclass correlation coefficients (ICC). This coefficient indicates the degree to which a single mark on a response represents the other marks on the same response (Shrout and Fleiss 1979). In general, an ICC value between 0.75 and 0.90 is considered good reliability and a value above 0.90 indicates excellent reliability (Cicchetti 1994). The ICC values of each

Figure 6. Histograms of examiner and auto-marker raw scores without hybrid marking



test part as well as the whole test are presented in Table 2. It can be seen that the reliability of single human marking varies from 0.84 to 0.91 in the five test parts and is 0.91 for the whole test, thus indicating adequate reliability of human marking at task level and excellent reliability at the test level. Brenchley (2020) re-examined inter-rater reliability using a larger dataset of 204 Linguaskill Speaking tests marked independently by three examiners. The study reported a single-marker ICC of 0.90 for the whole test.

6.2.2 Reliability of the auto-marker on its own

Auto-marker evaluation is often performed by computing the correlation or agreement between computer marking and human marking (e.g., Bernstein et al 2010, Wang et al 2018). Jones, Brenchley and Benjamin (2020) conducted an evaluation study on the current version of the auto-marker, focusing first on the performance of the auto-marker on its own – that is, not embedded in a hybrid marking system (see next section). The evaluation was based on a dataset of 9,286 Linguaskill Speaking tests reflecting live candidature. The distribution of human CEFR grades in the dataset was approximately 1% Below A1, 5% A1, 13% A2, 36% B1, 35% B2 and 10% C1 or above. The dataset contained speakers of a large number of native languages in which Spanish (29%), Arabic (26%) and Portuguese (15%) were the most frequent.

The study found that when the same CEFR cut-off values are applied to auto-marker and human raw scores, the auto-marker awarded the same CEFR grade as the examiners in 56.8% of the tests. In 96.6% of the tests, the difference between computer marking and human marking was equal to or smaller than one CEFR level. In 3.4% of the tests, the auto-marker and human examiners differed by more than one CEFR level (Table 3). In operational testing, these inaccurate auto-marker scores are overridden by examiner scores, which will be discussed in the next section.

The research also found that although the distributions of the auto-marker and human raw scores largely overlapped (Figure 6), the auto-marker was comparatively harsher on the higher end of the scoring scale and comparatively lenient on the lower end (Figure 7). Again, these inaccuracies are addressed through the use of hybrid marking, as well as continued training and improvement of the auto-marker. The root mean square error (RMSE) of the auto-marker raw score was 0.64, about a half CEFR band. RMSE is the standard deviation of the residuals (auto-marker prediction errors) and an indicator of how concentrated the data points are around the diagonal regression line where exact human–machine agreement is achieved (Figure 7).

In addition to auto-marker agreement with examiners, the research also evaluated the usefulness of the Language Quality score, a confidence measure of speech recognition (see Section 4.3), for identifying non-English speech or gibberish in the test responses. Based on a subset of data (n = 284) which included aberrant speaking behaviours, it was found that normal English speech resulted in significantly higher Language Quality scores than non-English speech (see Figure 8). By applying a cut-off value to this score,

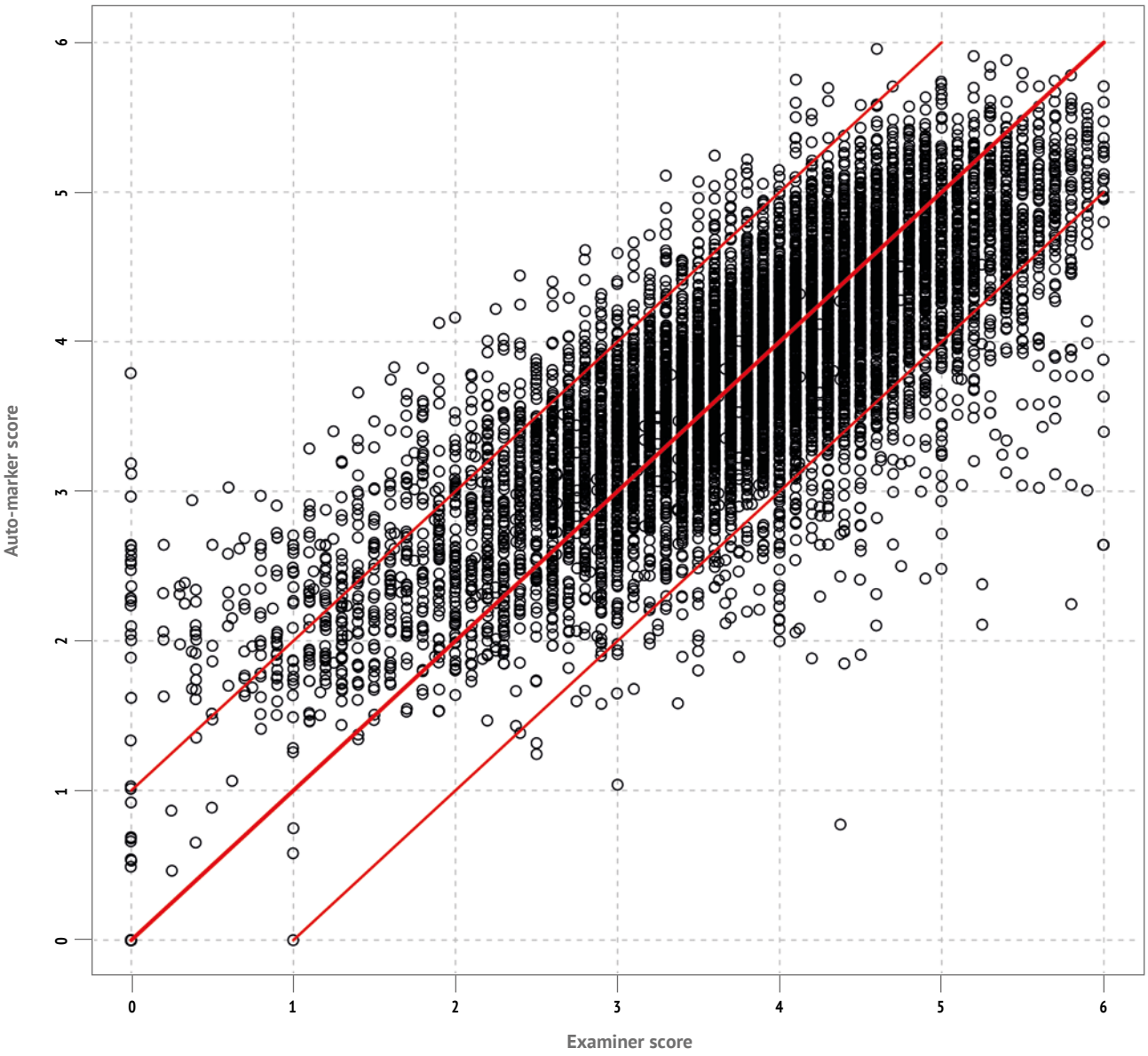
Table 2. Intraclass correlation coefficients of single examiner marking (Xu and Gallacher 2017)

Part 1	Part 2	Part 3	Part 4	Part 5	Whole test
0.84	0.87	0.90	0.88	0.91	0.91

Table 3. Percentage agreement between auto-marker and human CEFR grades (n = 9,286)

Human–machine agreement	Percentage
Exact agreement (or no difference)	56.8%
Adjacent agreement (or difference <= 1 CEFR level)	96.6%
Mismarking (or difference > 1 CEFR level)	3.4%

Figure 7. A scatter plot of examiner vs. auto-marker raw scores without hybrid marking



the 19 non-English-speaking responses in the dataset were all successfully identified. The research suggests that the Language Quality score is sensitive to non-English speech and helpful for recognising candidates with an intent to game the auto-marker.

6.2.3 Reliability of hybrid marking

Hybrid marking, as mentioned in Section 4.3, is about escalating to human examiners any responses which the auto-marker may have *mismarked*, defined as cases where the auto-marker and human scores are likely to be further than one CEFR level apart on the scoring scale. In the Linguaskill hybrid marking model, rules are applied to a number of features generated by the auto-marker including Assessment Quality, Language Quality, Audio Quality, auto-marker score (lower bound) and auto-marker score (higher bound). Each rule is an inequality statement such as *the Language Quality score is below 0.9*. If a response satisfies any of the rules, it is passed to a human examiner. The thresholds that are used in the rules (e.g. 0.9) were determined by a process of optimisation.

This constrained optimisation was done using exhaustive search, also known as brute-force search. For each variable a set of possible thresholds was created. For example, the Language Quality score ranges from 0 to 1, so the set of thresholds might be 0, 0.01, 0.02, ..., 1. For all possible combinations of five thresholds, an optimal (highest) recall statistic was calculated based on the dataset of 9,286 Linguaskill Speaking tests (Jones et al 2020). The recall statistic, which is reported as a percentage, indicates the completeness of flagging. For example, a recall value of 0.90 means that of all the test responses mismarked by the auto-marker, 90% of them are successfully flagged by the application of the rules.

Figure 8. Language Quality scores between English-speaking tests and non-English-speaking tests

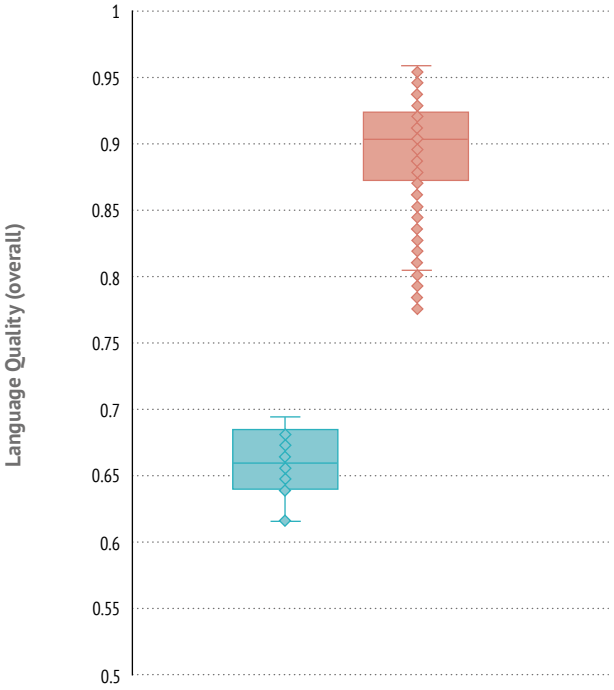
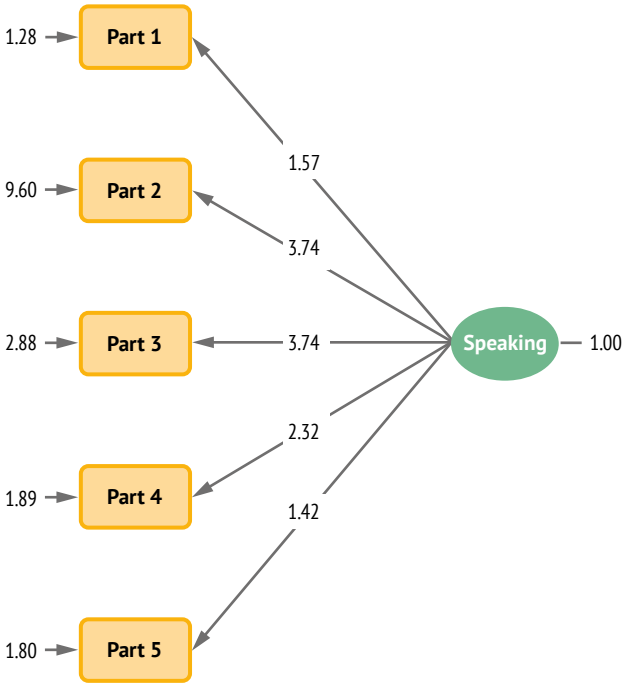


Figure 9. Single Factor model (Xu and Seed 2017)



A statistic that is often reported along with recall is precision, which is an indicator about the accuracy of flagging. For example, a precision of 0.90 means that of all the flagged test responses, 90% of them were indeed mismarked by the auto-marker. There is always a trade-off between precision and recall – a high recall value will lead to a low precision value and vice versa. In designing the Linguaskill piping rules, we pursued a high recall value in order to prevent unreliable auto-marker scores being released to the candidates.

Given our emphasis on high reliability of marking, we initially opted for threshold values that would result in a recall of 0.96 at the cost of escalating a large proportion of test responses to human examiners. The high recall, in turn, led to a small auto-marker RMSE of 0.16 and excellent human–machine agreement: 95.6% exact agreement and 100% adjacent agreement on CEFR grades. We are, however, continually improving the auto-marker and evaluating the threshold values to decrease the proportion of test responses that are examiner-marked.

6.3 Validity evidence for Interpretation of Test Results

The interpretation of language test results must be supported by construct theories about targeted language abilities. On the one hand, construct theories about language are chosen by test developers to inform test design, assign meanings to the test scores and account for the variance in test scores. On the other hand, test validation is also a process of theory validation in that the observed test data may either confirm or refute the chosen theories for score interpretation (Cronbach and Meehl 1955).

A construct theory may be a set of language proficiency descriptors, as in the CEFR, which detail the course of language development. Alternatively, it can be a speculation on the composition of a language ability. The validity evidence for supporting the proposed score interpretation of the Linguaskill Speaking test has been collected via standard setting and factor analysis. The former links the performance on the test to a theory about speaking proficiency progression whereas the latter examines the underlying structure of the speaking construct targeted by the test.

6.3.1 Standard setting

As the Linguaskill Speaking test reports CEFR-based test results, standard-setting exercises were performed periodically to align its test results to the CEFR framework. This alignment allows test users to interpret the test results in a wider context by referring to the language proficiency descriptors provided by the CEFR.

Standard setting refers to *the process of establishing one or more cut scores* on examinations (Cizek and Bunch 2007, p. 13). In the case of Linguaskill, cut scores are used to divide candidates into six proficiency groups in line with the CEFR proficiency levels: Below A1, A1, A2, B1, B2 and C1 or above. The most recent standard-setting exercise on the Linguaskill Speaking test was conducted by Lopes and Cheung (2020) who followed a modified Bookmark method recommended by a manual for relating language tests to the CEFR (Council of Europe 2009).

6.3.2 Factor structure

In addition to standard setting, factor analysis was performed to examine the underlying structure of the Linguaskill Speaking test. It was hypothesised that the abilities assessed

in the five test parts were unidimensional, meaning that a single, overarching speaking construct was assessed by the test. However, it appeared that Reading Aloud, the second part of the test, might assess a slightly different construct from the other four spontaneous speaking tasks.

To test the above hypothesis, Xu and Seed (2017) conducted an item-level confirmatory factor analysis on 3,250 speaking tests solely marked by examiners. The study found that a Single Factor model (Figure 9) fit the data well, resulting in a Comparative Fit Index (CFI) value of 0.99, a Non-Normed Fit Index (NNFI) value of 0.98, and a Root Mean Square Error of Approximation (RMSEA) value of 0.08. Generally, a CFI or NNFI value of 0.90 or above or an RMSEA value of 0.80 or below indicates an adequate model fit (Sawaki et al 2009). The finding suggests that a single speaking construct was able to account for test performances in all the five parts, thus supporting the practice of averaging the five parts to produce an overall test score. It was, however, also noted that the residual (error) term associated with Part 2, Reading Aloud, was relatively larger than those associated with the other parts. The researchers regarded this as a piece of evidence for distinguishing between reading aloud and spontaneous speaking in speaking assessment, and cautioned against using constrained speaking tasks alone to measure communicative speaking ability.

References

AERA, APA and NCME (2014) *Standards for educational and psychological testing*, Washington, DC: AERA.

APA, AERA and NCME (1974) *Standards for educational and psychological tests*, Washington, DC: APA.

Bachman, L F and Palmer, A S (1996) *Language testing in practice*, Oxford: Oxford University Press.

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

Bernstein, J, Van Moere, A and Cheng, J (2010) Validating automated speaking tests, *Language Testing* 27 (3), 355–377.

Brenchley, M (2020) *Re-examining the reliability of human marking in the Linguaskill Speaking test*, Cambridge Assessment English internal research report.

Brown, A (2003) Interviewer variation and the co-construction of speaking proficiency, *Language Testing* 20 (1), 1–25.

Cambridge Assessment English (2016) *Principles of good practice: Research and innovation in language learning and assessment*, Cambridge, UK: Cambridge Assessment.

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

Cheng, L (2014) Consequences, impact, and washback, in Kunnan, A J (Ed.) *The Companion To Language Assessment* (Vol. III) Chichester, West Sussex: John Wiley and Sons, 1,130–1,146.

Chun, C W (2006) An analysis of a language test for employment: The authenticity of the PhonePass test, *Language Assessment Quarterly* 3 (3), 295–306.

Cicchetti, D V (1994) Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology, *Psychological Assessment* 6 (4), 284–290.

Cizek, G J and Bunch, M B (2007) *Standard setting: A guide to establishing and evaluating performance standards on tests*, Thousand Oaks, CA: Sage.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, teaching, assessment*, Strasbourg: Council of Europe.

Council of Europe (2009) *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A Manual*, Strasbourg: Council of Europe.

Council of Europe (2018) *Common European Framework of Reference for Languages: Learning, teaching, assessment (Companion volume with new descriptors)*, Strasbourg: Council of Europe.

Cronbach, L J (1988) Five perspectives on validity argument, in Wainer, H and Braun, H I (Eds) *Test validity*, Hillsdale, NJ: Lawrence Erlbaum, 3–17.

Cronbach, L J and Meehl, P E (1955) Construct validity in psychological tests, *Psychological Bulletin* 52 (4), 281–302.

Fan, J (2014) Chinese test takers' attitudes towards the Versant English Test: A mixed-methods approach, *Language Testing in Asia* 4 (6), 1–17.

Galaczi, E and Taylor, L (2018) Interactional competence: Conceptualisations, operationalisations, and outstanding questions, *Language Assessment Quarterly* 15 (3), 219–236.

Haertel, E H (2006) Reliability, in Brennan, R L (Ed.) *Educational Measurement* (4th edn), Westport, CT: Praeger, 65–110.

Jones, E, Brenchley, M and Benjamin, T (2020) *An investigation into the hybrid marking model for the Linguaskill Speaking test*, Cambridge Assessment English internal research report.

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

Knill, K, Gales, M, Kyriakopoulos, K, Malinin, A, Ragni, A, Wang, Y and Caines, A (2018) Impact of ASR Performance on Free Speaking Language Assessment, *Proc. Interspeech* 2018, 1,641–1,645. https://doi.org/10.21437/Interspeech.2018-1312

Linacre, J M (1989) *Many-facet Rasch measurement*, Chicago: MESA Press.

Lopes, S and Cheung, K (2020) *Final report on the December 2018 standard setting of the Linguaskill General papers to the CEFR*, Cambridge Assessment English internal research report.

Lu, Y, Gales, M, Knill, K, Manakul, P, Wang, L and Wang, Y (2019) Impact of ASR performance on spoken grammatical error detection, *Proceedings of the Annual Conference of the International Speech Communication Association*, *INTERSPEECH*, September 2019, 1,876–1,880. https://doi.org/10.21437/Interspeech.2019-1706

Malinin, A, Ragni A, Knill, K and Gales, M (2017) Incorporating Uncertainty into Deep Learning for spoken language assessment. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* 2, 45–50. https://doi.org/10.18653/v1/P17-2008

Messick, S (1989) Validity, in Linn, R L (Ed.), Educational measurement (3rd edn), New York: Macmillan, 13–103.

Messick, S (1996) Validity and washback in language testing, *Language Testing* 13 (3), 241–256.

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

Shrout, P E and Fleiss, J L (1979) Intraclass correlations: Uses in assessing rater reliability, *Psychological Bulletin* 86 (2), 420–428.

van Dalen, R C, Knill, K and Gales, M (2015) Automatically grading learners' English using a Gaussian process. *SLaTE 2015: Workshop on Speech and Language Technology in Education*, 7–12. https://www.isca-speech.org/archive/slate_2015/sl15_007.html

Wang, Y, Gales, M J F, Knill, K M, Kyriakopoulos, K, Malinin, van Dalen, R C and Rashid, M (2018) Towards automatic assessment of spontaneous spoken English, *Speech Communication* 104, 47–56. https://doi.org/10.1016/j.specom.2018.09.002

Weir, C J (2005) *Language testing and validation: An evidence-based approach*, Basingstoke: Palgrave Macmillan.

Xi, X (2010) Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing* 27 (3), 291–300.

Xi, X, Schmidgall, J and Wang, Y (2016) Chinese users' perceptions of the use of automated scoring for a speaking practice test, in Yu, G and Jin, Y (Eds) *Assessing Chinese learners of English: Language constructs, consequences and conundrums*, Basingstoke, Hampshire: Palgrave Macmillan, 150–175.

Xu, J (2015) *Predicting ESL learners' oral proficiency by measuring the collocations in their spontaneous speech*, unpublished doctoral dissertation, Iowa State University, Ames, IA.

Xu, J and Gallacher, T (2017) *Linguaskill Speaking trial report*, Cambridge Assessment English internal research report.

Xu, J and Seed, G (2017) *Automated speaking tests: Merging technology, assessment and customer needs*, paper presented at the Language Testing Forum 2017, Huddersfield, UK.

Contact us

We are Cambridge Assessment English. Part of the University of Cambridge, we help millions of people learn English and prove their skills to the world.

For us, learning English is more than just exams and grades. It's about having the confidence to communicate and access a lifetime of enriching experiences and opportunities.

With the right support, learning a language is an exhilarating journey. We're with you every step of the way.

Cambridge Assessment English
The Triangle Building
Shaftesbury Road
Cambridge
CB2 8EA
United Kingdom

cambridgeenglish.org
/cambridgeenglish
/cambridgeenglishtv
/cambridgeeng
/cambridgeenglish
/cambridge-assessment-english

All details are correct at the time of going to print in June 2020.